School accountability and fairness: Does 'Progress 8' encourage schools to work more equitably?

Simon Burgess

Dave Thomson

December 2020

Acknowledgments

The Nuffield Foundation is an independent charitable trust with a mission to advance social well-being. It funds research that informs social policy, primarily in Education, Welfare, and Justice. It also funds student programmes that provide opportunities for young people to develop skills in quantitative and scientific methods. The Nuffield Foundation is the founder and co-funder of the Nuffield Council on Bioethics and the Ada Lovelace Institute. The Foundation has funded this project, but the views expressed are those of the authors and not necessarily the Foundation. Visit www.nuffieldfoundation.org

We are also grateful to the Department for Education for access to the National Pupil Database and to members of our Advisory Group for comments on earlier versions of this report. Any errors are ours.

Contents

Executive Summary

Background

School accountability in the form of published "league tables" of performance indicators and routine inspection have been a feature of the education system in England since the early 1990s. This creates incentives, both positive and negative, to which schools respond. For example, changes to the list of qualifications that count towards a school's published performance indicators will change the nature of qualifications offered by schools.

In 2013, the UK Government announced an important reform to the accountability framework for state-funded schools in England by publishing a new performance indicator: Progress 8. This change was notable as the previous framework was dominated by a threshold measure of pupil attainment, whether or not a pupil achieved 5 or more General Certificates of Secondary Education (GCSEs) at grades A*-C including English and maths (5ACEM).

The reform attempted to address two concerns. Firstly, that the 5ACEM indicator encouraged schools to focus on a particular set of pupils, namely those that were on the grade C/D borderline. Secondly, it was recognised that published statistics on secondary schools' attainment did not account for differences in attainment on entry to school.

Progress 8 is a value-added measure, summarising attainment at the end of compulsory secondary education (usually at age 16), controlling for attainment at the end of primary education (usually at age 11). Unlike the previous regime, it offered no particular incentive for schools to focus on a narrow segment of the pupil population. Not only that, it was felt to offer a fairer way of comparing the performance of schools. Although measures of value-added had been published for many years, they had not been conferred with the same level of precedence as Progress 8.

Research aims and methodology

We aimed to understand whether the introduction of Progress 8 encouraged schools to work more equitably. By this, we mean whether we find evidence of schools focusing their efforts less on pupils at the C/D borderline and instead spread their effort more evenly across the full range of attainment.

Administrative data on the attainment of all pupils in state-funded schools in England is available dating back to 2001/02. We attempt to use such data to estimate the effect of the introduction of Progress 8 on the attainment of above-borderline and below-borderline pupils relative to borderline pupils.

However, there are a number of methodological issues to overcome. Firstly, we do not know which pupils schools considered to be at the C/D borderline. Secondly, Progress 8 was just one of a series of policy changes that have occurred to the secondary school accountability framework in England since 2010. Thirdly, improvements in examination results from year to year are controlled by the regulator for qualifications, Ofqual, by a method known as comparable outcomes.

Our analytical approach is shaped by these issues. We therefore use six years of pupil-level data covering all state-funded school pupils in England. This starts in 2011/12, the first year that the Comparable Outcomes policy was applied to GCSE results in English and maths. Progress 8 was first published in 2015/16. This means we have data for four years prior to its introduction and two years since its introduction. We also track changes in pupil attainment using a number of indicators that are reasonably stable in definition and coverage over the period we observe. We also observe changes in the types of qualifications pupils entered.

We adopt a difference-in-difference (d-in-d) approach to isolate the causal effect of the reform on a number of pupil outcomes. That is, we compare outcomes between two groups of pupils ("difference") before and after the policy change ("difference in difference"). We examine, following the reform, changes in the outcomes for the different groups of pupils that theory suggests will be differentially affected. We use a flexible approach to modelling that controls for a wide range of pupil characteristics, and for school factors and other policy shocks.

Key findings

Our results are consistent with the view that some schools had reacted to the previous regime of high implicit incentives for the exam results of students at the GCSE grade C/D borderline. Once that incentive was removed, the borderline group appeared to make less relative progress compared to other groups. The effects are small but not trivial: our headline findings show a post-reform gain of 0.01 standard deviations (SD) in GCSE English and maths for the above-borderline group and 0.06SD for the below-borderline group. This latter effect on GCSE attainment is the same size as that arising from a 1SD increase in school expenditure (Jenkins et al, 2005).

We are, however, cautious in presenting these results noting the issue of trends subsequent to announcement but before implementation. We judge the results to be supportive of the hypothesis but not clinching. We are also aware that the results may have been sensitive to some of our modelling choices. Therefore, we show the effects of making different choices. Our tests of robustness show that these different choices make little substantive difference.

The results also have a bearing on the test score gap between disadvantaged pupils and their peers. Because disadvantaged pupils are disproportionately likely to be in the below-borderline group, and so are more likely to gain from the reform, our findings show a slight post-reform improvement of 0.01 SD.

The change to the accountability framework was far-reaching and had other implications beyond simply test scores. The machinery of school accountability also incentivises schools to enter pupils for particular qualifications. Just as the Government response to the Wolf Review had done two years earlier, the introduction of Progress 8 led to large changes in the types of qualification for which pupils were entered as schools increasingly began to fill the eight qualification 'slots' available in the new accountability measures. In most cases, this was a result of switching away from other types of qualification that were not eligible for inclusion in the accountability measures.

Changes in school behaviour

Schools' responses to the introduction of Progress 8 were varied, although it is difficult to disentangle specific responses to Progress 8 from responses to other changes that happened around the same time, such as reforms to GCSEs. We surveyed over 400 school leaders and teachers in England to find out more about how they responded to the introduction of Progress 8. The results suggest a general shift away from running intervention sessions aimed specifically at borderline pupils towards pupils judged to be falling behind.

Policy Implications

This analysis has a number of implications for policy-makers.

First, and bearing in mind the caveats noted in the report, our results suggest that the introduction of Progress 8 had the intended effect of shifting schools' focus away from students who were borderline to the previous accountability threshold. In that sense, the policy had the intended effect of making schools work more equitably.

Second, this reinforces the view that accountability measures are an effective policy tool. They do not impinge directly on schools' operational autonomy, unlike explicit Ministerial directives, but they do adjust the incentive structure that schools face. This research shows that this can be effective in changing behaviour. The setting, and occasional re-setting, of the accountability framework seems an appropriate role for Government – it is the practical expression of its view of what society deems valuable in education, of what schools 'ought' to do. Problems arise if the framework is changed very frequently so that schools do not have a stable environment for planning.

Problems can also arise if different parts of schools' incentives pull in different directions, and this is the third and final policy message from this study. The previous accountability regime was based on the threshold of achieving 5 or more GCSEs at grades A*-C including English and maths (5ACEM). Schools were strongly incentivised to maximise the number of their pupils that achieved this. This drive meshed well with the goal of the typical pupil because for her, passing that threshold was the key to accessing higher or further education and to the job market. Schools could allocate their resources knowing that the goal of doing well by their pupils and the goal of doing well on the performance metrics were reasonably well aligned. In the new regime, currently, that remains true for pupils but less so for schools. Access to further education and to jobs is still to an extent dominated by achieving at least grade C (now grade 4) passes in GCSE English and maths, and no attention is paid to the achievement of pupils in Progress 8 terms by employers. This may mean that schools are partially conflicted, and that a goal for the school of keeping the 5ACEM "pass rate" high is still important to them. This in turn may partly explain why the impact of the reform on test scores was rather modest. It may be that the labour market and higher education admissions departments will respond and place more emphasis on Progress 8 scores, or it may be that these two goals for schools will remain in tension.

1 Introduction

Schools have a key role in generating the educational outcomes that are central for a country's prosperity, inequality and social mobility. For some time, researchers have been interested in the incentives and constraints facing schools, considering that they may be a tool to influence school performance. One of these incentives comes from school accountability. The use of examination data by the Government to hold schools to account has been a feature of the education system in England since the early 1990s. Each year, Performance tables are published which, in theory, encourage schools to improve their performance, thereby attracting pupils and funding. In addition, the threat of sanctions has tended to face those schools falling below the minimum performance thresholds known as floor standards (formerly known as floor targets) although this threat has been relaxed somewhat in recent years (Department for Education, 2019a).

The impact of high stakes accountability has been contested in previous research. Burgess et al (2013) used the abolition of the school performance tables in Wales after devolution as an exogenous shock to the accountability regime there. They show that pupil test scores deteriorated significantly afterwards relative to those of equivalent pupils in England. They showed that this deterioration was particularly marked for lower ability pupils and pupils from disadvantaged areas. Others, for example Foley & Goldstein (2012), highlight the limitations of the league tables and the implications for accountability.

This research takes the analysis an important step forward by focussing specifically on the form of the accountability mechanism. We exploit a recent policy change to the accountability system in England to see how that has changed school behaviour and to evaluate any impact it has had on pupil attainment. The policy change is the introduction of a measure called 'Progress 8' as the headline accountability measure in 2015/16 (following a limited pilot and "test-run" in 2014/15).

While there have been many adjustments to the accountability system since the 1990s, this reform marks a radical departure. Progress 8 differs in two key ways from prior systems: it is based on a Value-Added (VA) approach rather than on raw outcomes, and second, it simply averages pupil outcomes for a school, rather than using a threshold approach. The first difference largely affects schools' admissions strategies because the use of raw outcomes in performance tables means that admitting high-performing pupils is helpful; a VA system reduces that incentive. The second difference is mostly about schools' decisions on teaching the pupils they have; this is the focus of this paper. A threshold measure, such as the fraction of pupils getting at least five good passes at GCSE, gives the school very strong incentives to raise the performance of pupils around the threshold, but much weaker incentives for pupils either way above the threshold (certain to get five good passes), or way below (almost zero chance) (see Wilson et al, 2004; Astle et al, 2011). Analysis of the incentives associated with this indicator showed that the regime reduced the exam performance of very low ability pupils (Burgess et al, 2005). A simple average across all pupils like Progress 8 does not present schools with any clear "priority" pupils. However, as with all performance measures, it has its own perverse incentives. In addition to evaluating the impact of the reform on attainment, we also examine the issue of pupils moving off the school roll, by permanent exclusion or otherwise, prior to taking examinations and therefore not being included in the measure.

We aim to identify the causal effect of the change in accountability on pupil achievement. We will interpret any change as resulting from changes in school behaviour, and report results from qualitative work (a questionnaire sent to secondary schools) to gain a better understanding of how behaviour changed in schools. While the policy change itself is exogenous, as is typically the case, it (unhelpfully) affects all schools at the same time.

We adopt a difference-in-difference (d-in-d) approach to isolate the causal effect of the reform on a number of pupil outcomes. We examine if, following the policy changes, the outcomes for different groups of pupils that theory suggests will be differentially affected. Although it is likely that schools responded to the incentives of the accountability system in different ways before and after the reform, we focus on the group of pupils most likely to be most affected by the reform, namely those who were marginal to the sharp threshold of five A*- C grades, the "borderline" group. Of course, the latter are not exogenous or unchanging features of schools or pupils. We deal with this in two ways. First, at a pupil level, we use the pre-reform definition of borderline (close to the C/D border) and carry that over into the post-reform period to ensure close comparability of groups. Second, at a school level, our analysis is limited to pupils who would have been in the school from before the policy change.

Structure of this report

We begin by setting out in Section 2 the relevant policy changes affecting school accountability in England since the early 1990s. Section 3 sets out our research questions and describes the data we use in our analysis, which is sourced from the National Pupil Database. In Section 4 we set out how we define the borderline group of pupils and our strategy for identifying the impact of the Progress 8. Section 5 presents the overall results of our quantitative analysis and tests of the robustness of our findings to alternative specifications. Results from our survey of schools are presented in Section 6. Finally, we offer some lessons for policymakers from this study.

2 Background

School accountability, in the form of published performance indicators and routine inspection, are an established yet often contested feature of the schools system in England, introduced following the Education (Schools) Act of 1992.

Over the intervening almost three decades, both the indicators published in School Performance tables and the framework used by the Office for Standards in Education (Ofsted) to inspect schools have been subject to change and revision. Although published data on school performance has always been a key source of evidence, inspection is meant to evaluate the work of schools more roundly, including safeguarding, governance, extra-curricular activity and so forth.

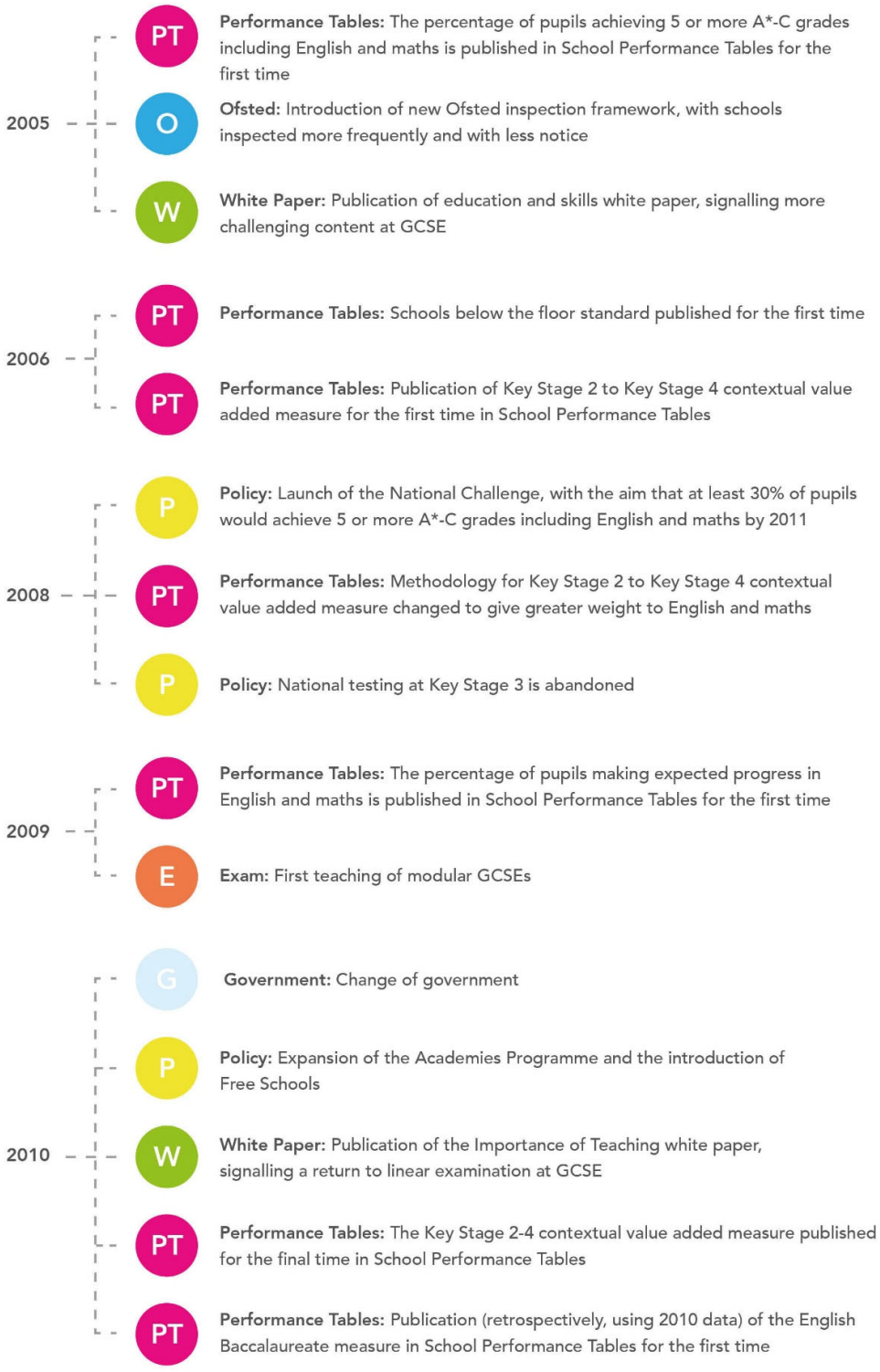
This report concerns itself with a single reform to the secondary school accountability regime in England, the introduction of Progress 8 as a headline measure of school performance in 2015/16. However, it was just one of a series of reforms to both the accountability regime and to the school system in England more generally over the last twenty to thirty years. These changes have a material bearing on our analysis, in which we try to isolate the effect of the introduction of Progress 8 from the effects of other changes occurring around the same time. We consider this in greater detail in Section 3.6.

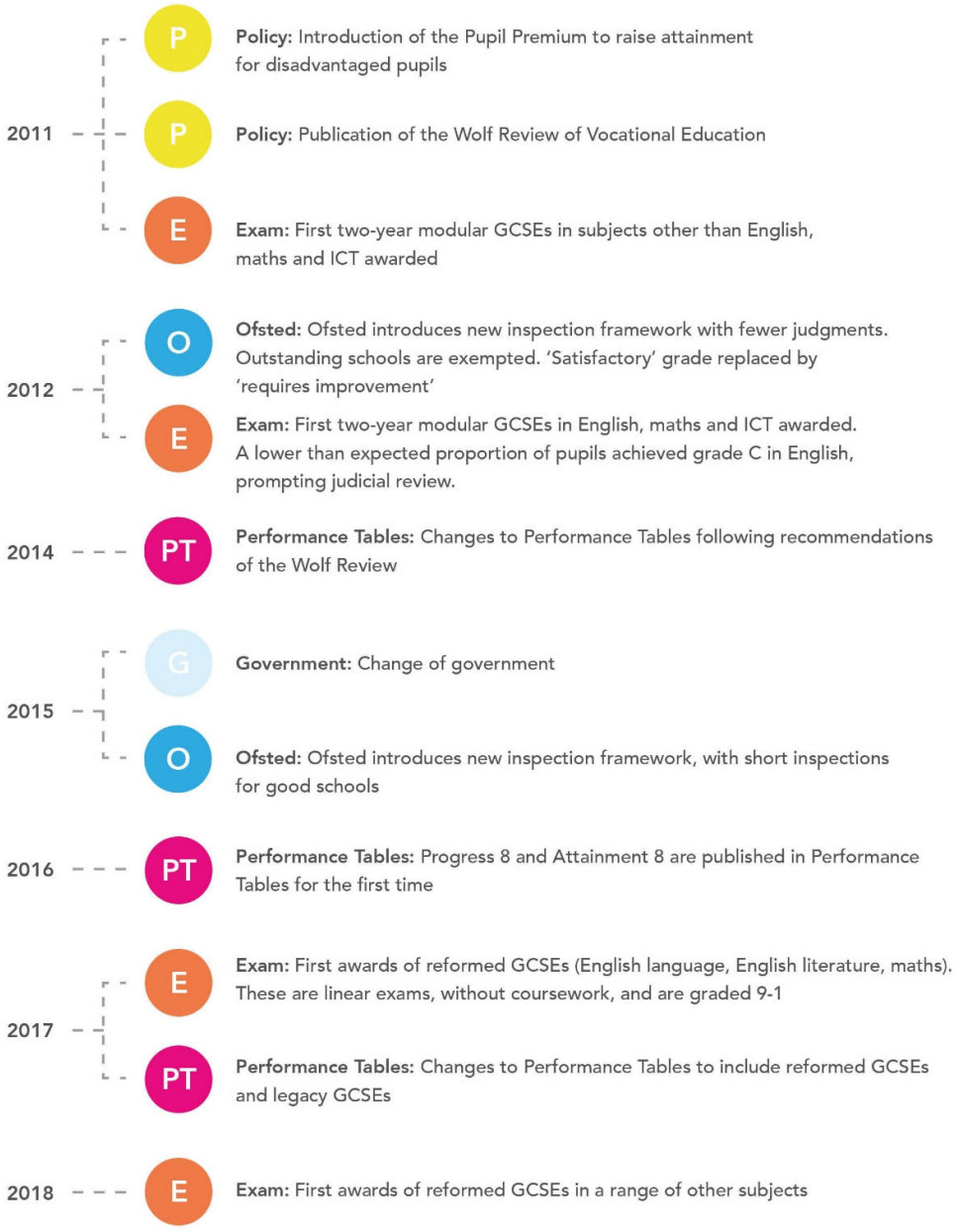
Running alongside changes to school accountability are changes to the assessments taken by young people during their compulsory schooling and to the qualifications taken at the end. The most common qualification, the General Certificate of Secondary Education (GCSE), has evolved since its introduction in 1988. Changes have seen a shift from linear examination to modular examination and back to linear examination, different approaches to coursework and controlled assessment and changing content within each subject (Ofqual, 2019).

Furthermore, the school system itself has undergone substantial reform this century with the introduction of Academies, a new type of autonomous state-funded school outside of local authority control and exempt from following the National Curriculum (Eyles and Machin, 2018). Under the New Labour government of 1997-2010, a relatively small number of these schools replaced low-performing state-schools. However, the Academies Programme was accelerated under the Coalition government of 2010-2015. Schools rated outstanding or good by Ofsted could convert to Academy status. At the start of our analysis period in 2004/05, there were just 11 secondary Academies. By 2015/16, the year Progress 8 was introduced, there were over 2,200; almost two-thirds of all state-funded secondary schools.

In this section, we present an overview of the evolution of secondary school performance indicators, the qualifications taken by 16 year olds, floor standards and performance incentives for schools and the introduction of Progress 8. The key changes are summarised in Figure 1.

Figure 1: Key accountability changes for secondary schools 2005 to 2018





2.1 A brief history of measuring secondary school performance

Secondary School Performance tables have been published every year since 1992. Their content and format has tended to change annually, with a statement of intent published in advance detailing changes for the forthcoming year (Department for Education, 2019b).

Historically, one indicator has tended to assume precedence although this has changed periodically. Initially, only achievements in GCSEs were counted and the percentage of

pupils achieving five or more A*-C grades became the headline measure of school attainment.

Performance tables were expanded in 1997 to include General National Vocational Qualifications (GNVQ). By 2003, pupils at the end of Key Stage 4 were entering some 135 thousand GNVQs. In order to produce school performance indicators, the equivalence of GNVQs to GCSEs was established. A pass in a full intermediate GNVQ was considered to be equivalent to four A*-C passes in GCSEs.

In 2000, a comprehensive state school in Shropshire, Thomas Telford, achieved an unparalleled 100% on the headline five A*-C performance indicator. An article published the following year in *The Guardian* (Revell, 2001) revealed its secrets, including the use of full intermediate GNVQ and early entry in GCSE mathematics. This was an early example of what would pejoratively be called "perverse incentives", "gaming" or "strategic behaviour" (Goldstein and Foley, 2012; Muriel and Smith, 2011; Ingram et al, 2018) to describe the practice of schools maximising performance measures without necessarily improving teaching and learning.

The 2005 White Paper on 14-19 Education and Skills (Department for Education and Skills, 2005) committed to "toughening" the performance tables by publishing the percentage of pupils achieving five or more A*-C grades at GCSE (or equivalent) including GCSE English and maths from 2005/06 onwards. This became the dominant headline measure until its replacement by Progress 8 in 2015/16.

Throughout the period we undertake our analysis, from 2005/06 to 2016/17, proxy measures of progress have also been published (Leckie and Goldstein, 2017). These recognise that schools differ in intakes and that pupil prior attainment plays a large part in determining subsequent attainment. A measure of prior attainment covering the entire period of secondary education from the end of Key Stage 2 (age 11) to the end of Key Stage 4 (age 16) was first piloted in 2003. This was similar in nature to Progress 8 in that pupils were banded according to their mean Key Stage 2 prior attainment in English, maths and science and their Key Stage 4 attainment was compared to other pupils in the same prior attainment band[i]. Key Stage 4 attainment was measured by awarding points to GCSEs and equivalent qualifications based on grade achieved and summing the best eight for each pupil[ii]. This became known as a capped points score.

These very early (and simple) value added measures were then replaced by contextual value added measures (CVA) in 2007. This went much further and added further statistical controls for factors associated with Key Stage 4 attainment beyond prior attainment including age, gender, ethnicity, special educational needs, disadvantage (free school meal eligibility), pupil mobility and school composition (e.g. mean KS2 score of the cohort). Multilevel modelling was used to calculate CVA, with school scores shrunken towards the national mean based on the intra-class correlation and the number of pupils in the cohort.

CVA was subsequently abandoned after 2010 since the incoming coalition government believed it reinforced low expectations for some groups of pupils, particularly disadvantaged pupils (Department for Education, 2010). A value added measure was published in 2011 controlling solely for prior attainment although the multilevel models

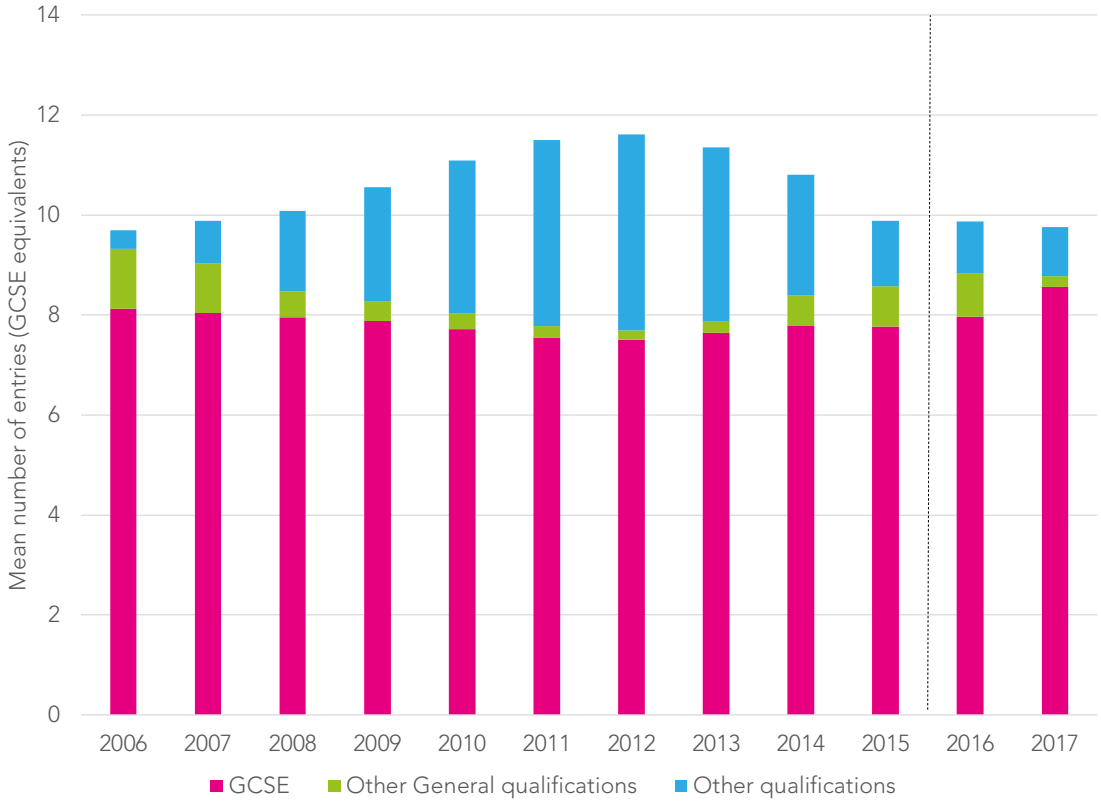
were retained. However, its importance was secondary in comparison to measures of "expected progress". First introduced in 2009, these were relatively simple indicators that summarised GCSE attainment in English (and maths) conditional on Key Stage 2 attainment in the same subjects. For instance, a pupil who achieved level 4 in English at Key Stage 2 would be deemed to have made expected progress if they achieved GCSE grade C (or above) in English. Unfortunately, these measures were biased with respect to prior attainment (Treadaway, 2015; Leckie and Goldstein, 2017). Pupils with higher levels of prior attainment were more likely to make expected progress.

2.2 A brief history of qualifications for 16 year olds

In 2004, the scope of performance tables was extended to include qualifications in addition to GCSE and GNVQ that were approved for use pre-16 under Section 96 of the Learning and Skills Act 2000. Just like GNVQs before them, these so-called "Section 96" qualifications would have their equivalence to GCSE established by the Qualifications and Curriculum Authority (QCA). At the same time, the National Curriculum was revised and modern foreign languages (MFL) and Design and Technology (DT) were made no longer compulsory at Key Stage 4.

In 2006, pupils entered an average of 9.7 qualifications measured in GCSE equivalents. Of these, 8.1 were in GCSEs, 1.2 were in other General qualifications (such as GNVQ) and the remaining 0.4 were in other (Section 96) qualifications. Data for other years is presented below (Figure 2). By 2012, pupils were entering an average of 11.6 qualifications, of which 3.9 were in Section 96 qualifications. The mean number of GCSEs entered had fallen to 7.5. Between 2006 and 2012, the percentage of pupils entering 12 or more GCSEs (or equivalents) rose from 17% to 48%.

Figure 2: Mean number of entries by qualification type, pupils at the end of Key Stage 4 in state-funded schools 2006 to 2017



Between 2008 and 2011, the practice of entering pupils more than once for GCSE English and mathematics started to become widespread. At its most extreme, some pupils were entered up to 12 times: in both winter and summer of Years 10 and 11, often with different awarding bodies (Ofsted, 2011). This would continue until checked by the Government's response to the Wolf Review of Vocational Education (Wolf, 2011).

From 2012 onwards, and following the Wolf Review and associated changes to School Performance tables, entries in Section 96 qualifications begin to decline. Most significantly, a raft of qualifications would no longer be counted, although would remain approved for young people under the age of 16. These included, among thousands of others, the hitherto popular literacy and numeracy skills qualifications and the short course GCSE in religious studies (Burgess and Thomson, 2019).

Several other changes addressed some of the prevailing Performance Tables incentives. Firstly, no single qualification would count as more than one GCSE. Secondly, a maximum of two non-GCSEs per pupil would be counted (this would later be increased to three when Progress 8 was introduced). Finally, the practice of entering a pupil more than once for the same qualification was checked by the phased introduction of a rule to count their first, rather than their best, result.

In our analysis of the impact of the Wolf Review (Burgess and Thomson, 2019), we found considerable variation between schools in the extent to which they entered pupils for non-GCSE qualifications. Although this practice was correlated with some school characteristics, particularly having lower levels of attainment in the recent past, schools with similar characteristics still tended to differ in the qualifications they offered. Similarly, although pupils entered for non-GCSE qualifications tended to have lower levels of Key Stage 2 attainment and be disadvantaged, there was still much variation in qualifications entered between pupils.

More reform to school accountability took place in 2015/16 with the introduction of Attainment 8 and Progress 8 (see below) and reforms to GCSEs. Over the course of three years, all subjects at GCSE would be reformed, with English and maths at the head of queue. This meant assessment by linear examination with other assessments being used rarely to test essential skills, more demanding content and a new grading scale running from 9 to 1. The first awards of reformed GCSE, in English and maths, were in 2016/17, the final year of data included in our analysis.

We show in Section 3.4.3 below further data on how qualification entry patterns changed with the introduction of Progress 8.

2.3 Comparable outcomes

Comparable outcomes is an approach taken by the exams regulator, Ofqual, to maintain grades awarded in public examinations by awarding organisations from year to year.

In brief, it refers to the use of statistics alongside the judgement of examiners and any other available evidence to ensure that standards are comparable from one year to the next. All things being equal, based on information on prior performance, the proportion of students who obtained certain grades this year should be the same as the proportion who achieved them last year (Baird et al, 2019). It recognises that judgments based on qualitative evidence alone can lead to grade inflation given the clustering of scores either side of a grade boundary (Cresswell, 1996).

The approach also limits the 'sawtooth' effect (Ofqual,2016) when new specifications are introduced, i.e. it does not unfairly penalise the first cohort of students who might otherwise achieve poorer grades as a result of teachers being unfamiliar with the new content.

Although first officially used in GCSE awarding in 2011 (Benton, 2016), its application first came into sharp relief when the percentage of pupils achieving A*-C in English fell in 2012 following increases every year since the introduction of GCSE in 1988. Ofqual was taken to judicial review by an alliance of pupils, schools, councils and professional bodies alleging that up to 10 thousand young people had inappropriately missed out on a grade C. Although it was ruled that Ofqual acted lawfully, the matter was subsequently reviewed by parliamentary committee (Education Committee, 2013).

Technical detail on the use of statistical data in the awarding process can be found in Benton and Sutch (2014).

2.4 Floor standards and performance incentives for schools

England operates a limited system of school choice in which parents have the right to express a *preference* for a particular secondary school. The publication of school performance data supports the choice-making process and so it is in schools' interests to be high attaining in order to attract sufficient pupils (Burgess et al, 2019).

However, the system is itself constrained by the number of places available. Guidance from Ofsted on the inspection of local authorities in the early part of the century suggested that there should be no more than 10% surplus places across the authority as a whole (Osborne, 2002) and no school should have more than 25% surplus places. This meant that pupils could be allocated to less popular schools, including those for which no preference had been made.

Performance data was also used heavily in Ofsted inspection during the period over which we conduct our analysis. Since 2005, schools have been graded on a four point scale (outstanding/good/ satisfactory/inadequate). The 'satisfactory' outcome was replaced by 'requires improvement' following revision to the inspection framework. Ofsted guidance from 2008 (Ofsted, 2008) claimed that performance data forms just part of the evidence for the inspection and is carefully evaluated. Later guidance, however, suggests a more mechanistic approach to analysis. Under the 2009 framework, a school could only be judged outstanding for attainment if over two-thirds of its performance indicators were (statistically) significantly above the national average (Ofsted, 2009).

Floor targets, later renamed floor standards, have also been used to incentivise schools to raise attainment. Announced in 2000, their purpose was to support schools whose performance fell below a specified 'floor', in this case 25% (later raised to 30%) of pupils achieving five or more A*-C grades (or equivalent). The bar was raised in 2008 to include A*-C passes in English and maths, with schools (and local authorities) being given additional support from the government's National Challenge initiative (Department for Children, Schools and Families, 2009). It was raised again in 2011, this time to 35% of pupils achieving the benchmark of five or more A*-C grades at GCSE (or equivalent) including English and maths (5ACEM). However, schools at which the percentage of pupils making expected progress in English or maths was above the national median were deemed to be above the floor. In 2012, the bar was raised further still, to 40%.

Although expected progress measures were used in defining the floor standards, they were largely determined by 5ACEM pass rates. For example, in 2014 (UK Government, 2019) fewer than 40% of pupils achieved 5ACEM at 401 out of more than 3,000 schools but just 44 were deemed to be above the floor standard when expected progress rates were considered.

The 2010 White Paper *The Importance of Teaching* (Department for Education, 2010) signalled a change in the purpose of floor targets. By now, the term "floor standards" was being used. Rather than leading to further support, the consequences were potentially more punitive. Most recently, these included warning notices from the Office of the Regional Schools Commissioner or forced academisation, an order from the Secretary of State for Education to become a sponsored academy.

Further pressure was exerted from the summer of 2016 through the identification of *coasting* schools by Michael Gove's successor, Nicky Morgan (Department for Education, 2015c). Whereas floor standards were determined on a single year's performance, coasting schools were identified according to performance in each of the last three years. The first coasting standard was based on the 5ACEM and expected progress indicators for 2014 and 2015 and the Progress 8 measure for 2016. Schools classified as coasting were required to produce a "clear plan for improvement" or face intervention from Regional Schools Commissioners.

Floor standards and coasting standards were removed in 2019 by Damian Hinds (Department for Education, 2019), the successor to Morgan's successor, as part of a package of measures to simplify the use of data across the system. This included a new framework for Ofsted inspection, one which accorded a lesser place to the use of data (Ofsted, 2019).

2.5 The introduction of Progress 8

Following a consultation on accountability measures for secondary schools, the Department for Education (DfE) announced in October 2013 that a new set of 'headline' measures would be published in January 2017 summarising school performance in the 2015/16 academic year (Department for Education, 2013).

Chief among these were a measure, Attainment 8. Like the previous "best 8" capped points score measure, Attainment 8 worked by allocating points to grades achieved in qualifications but was more prescriptive about which qualifications could be counted. There were eight slots (sometimes called buckets) in total: one for English, one for maths, three for the Ebacc subjects (sciences, humanities[iii], modern and ancient languages) and three 'open' slots for any other eligible qualifications. English and maths were double-weighted so there were effectively 10 slots in total.

Progress 8 was introduced as a new headline[iv] accountability measure. As the name suggests, it is closely related to Attainment 8. It differs in three ways from the predecessor headline measure based on the percentage of pupils gaining at least five A*-C grades. First, it inherits the structure of Attainment 8, and so involves the much stronger prescription of which subjects 'count'. Second, it is a value-added type measure, taking into account each pupil's prior attainment when they joined the school. The methodology is a simple one, banding pupils based on their mean Key Stage 2 fine grade in English and maths (Burgess and Thomson, 2013), later changed to reading and maths due to changes made five years earlier to Key Stage 2 assessments. But third, and most importantly, it is a simple average of all pupils in the school, not a threshold. This carries significant implications for schools' behaviour in terms of the removal of the borderline and schools' responses to it both before and after the introduction of Progress 8; this forms the basis of this report.

Schools could opt to participate in a pilot of Progress 8 in 2014/15. A total of 327 schools duly did so (Allen, 2015) with the added benefit that they would escape the 5ACEM floor standard if their score was above the suggested Progress 8 floor standard of -0.5.

Progress 8 would also be used to define a floor standard. In the consultation, the government had envisaged using a threshold measure in addition to a progress measure, namely the percentage of pupils achieving A*-C passes in the 'basics' of English and maths. However, this was dropped in its consultation response due to concerns that "there would be a continued incentive for schools to target teaching resources towards a small number of pupils close to a 'borderline' in English and mathematics" (Department for Education, 2013, pp 10).

3 Data and methods

In this section, we set out the aims of our research, the data we use and the methodology we employ to estimate the impact of the introduction of Progress 8 on a range of pupil outcomes.

3.1 Research aims and questions

Our main research aim is to advance our understanding of how school accountability systems influence school behaviour and therefore pupil outcomes. We know that schools respond to incentives, and the introduction of the Progress 8 system, a large systemic change, affects schools' incentives in relation to the accountability framework. Our central research question is to establish how, if at all, the introduction of the new accountability system built around Progress 8 has changed pupil outcomes. As noted, our hypothesis at the outset is that Progress 8 would encourage schools to switch their effort from pupils close to the C/D border towards higher and lower attaining pupils and that this shift will be observable in the data. Under some assumptions about the cost of raising attainment, we might expect to see no change. As the new accountability system rewards the overall average, it clearly does not matter where in the attainment distribution that arises. So if the costs of raising attainment are linear, and if there are switching costs for the school changing its practices, then we might see no response. However, if the costs of focussing more and more resources on one group are increasing at the margin, then it would pay the school to diversify away from that focus group.

We additionally compute the implications of these changes for pupils eligible for the Pupil Premium (PP), the additional funding stream introduced by the Coalition Government in 2011 to improve the attainment of disadvantaged pupils (West, 2015).

Of course, this is not a new idea. Reback (2008) used individual pupil-level data from the 1990s in Texas to study the effects of a school accountability system that was later transferred to the Federal level under the "No Child Left Behind" Act signed by President Bush in 2002. He finds that: "The empirical results suggest that schools respond to the accountability system by taking actions which influence the distribution of student achievement." Specifically, he shows that schools will target resources on students whose scores matter disproportionately for the overall accountability-relevant performance of the school. He concludes that "[i]f one of the primary goals is to create a sort of educational triage, in which students below minimum grade-level skills are pushed up, then the *No Child Left Behind* type of accountability system appears to be fairly effective.", but also notes that whether this is considered positive or negative depends on the welfare or policy weight on the specific group of students that are helped.

Schools responses to the changed incentives will depend to a degree on their context, on their 'market position'. A simple example is that selective (Grammar) schools, for instance, have very few (if any) pupils at the C/D border so the prior regime would have been irrelevant for them. We therefore then explore how different types of schools have responded to the change in headline accountability measure, including:

- Schools with different historical levels of pupil attainment and performance.

- Schools with different degrees of local competition
- Schools entering pupils for different portfolios of types of qualifications entered. Progress 8 encourages schools to enter pupils for particular qualifications (e.g. GCSEs in Ebacc subjects) which tend to be graded more severely than alternative non-GCSE qualifications.
- Schools with different proportions of borderline pupils

Note the first two of these might be thought exogenous, but the latter two are clearly chosen by the schools. For those results, the interpretation is different as we cannot claim they are necessarily causal.

A final aim is to examine the strengths and weaknesses of the methodological approach selected here. As an observational study, it will always rest on untestable assumptions. However, we can test the sensitivity of our results to different assumptions. The research strategy we propose here could be repeated in future years to assess the impact of further changes to the accountability regime, such as the introduction of 9-1 grades at GCSE.

3.2 Data

We use pupil-level administrative data from the National Pupil Database (NPD), maintained by the Department for Education, and described below. Our analysis period is GCSE exam results taken in 2012 (so at the end of the school year 2011/12), each year through the exams at the end of the school year 2016/17. Progress 8 was first published[v] in January 2017 relating to the exams taken in June 2016, so the 'after' years are 2015/16 and 2016/2017, and the 'before' years are 2011/2012, 2012/13, 2013/14 and 2014/15. However, equivalent data are available back to 2005/06, and we use this longer run to contextualise our period, and to justify our choice of analysis window.

The NPD houses datasets containing pupils' results in GCSEs and other approved qualifications at the end of compulsory schooling (Key Stage 4), usually at the age of 16. These records have been matched to details of prior attainment in tests and teacher assessments at the end of primary school (Key Stage 2), usually at age 11. Further data on pupils' school enrolments during their school career and personal characteristics (free school meal eligibility, ethnicity, special educational needs and so forth) is available from School Census, an annual collection between 2001/02 and 2004/05, becoming a termly collection thereafter. The principal dataset we use contains details of Key Stage 4 outcomes, personal characteristics and Key Stage 2 prior attainment for all pupils who reach the end of Key Stage 4 in state-funded mainstream schools between 2004/05 and 2016/17. These are the pupils who contribute to the measures published in the School Performance tables.

Pupils taking GCSEs in our two post-reform years, 2015/16 and 2016/2017, would typically have been in their schools well before the announcement of Progress 8 in October 2013, so the pupil-school match is exogenous to the policy. Pupils taking GCSEs in the summer of 2016 would typically have joined their school in 2011, and chosen their school in 2010, and pupils taking GCSEs one year later joined in 2012 and chosen in 2011.

However, some pupils would have changed school after the announcement of Progress 8. We use a different dataset to assess whether these changes affect our main results and to consider a second research question we consider, namely, whether Progress 8 encourages "off-rolling" pupils.

We therefore create a second dataset based on pupils we observe in Year 8, the second year of secondary education, attending state-funded mainstream schools between 2001/02 and 2012/13. These pupils would ordinarily have reached the end of Key Stage 4 between 2005/06 and 2016/17.

We use this dataset to test the robustness of our main results and to evaluate whether the introduction of Progress 8 has led to more instances of pupils leaving the school roll either by official exclusion or by being managed off-roll (Henshaw, 2017). In a small proportion of cases, which we discuss in section 3.4.2, we do not observe end of Key Stage 4 outcomes. Reasons for this include emigration, death, and moves into home education or the independent sector without any entries in approved qualifications.

These pupil-level datasets are supplemented with additional data about schools and their local areas. Further information about schools, including religious denomination, governance and admissions policy can be found in other administrative sources, such as *Get Information About School* (UK Government, 2020). Data on local neighbourhoods, such as the Income Deprivation Affecting Children Index (IDACI) can be added to pupil records as the lower super output area (LSOA) in which they reside are contained in NPD.

The qualifications pupils are observed to have entered, and the performance indicators calculated for them in NPD, are dependent on the prevailing accountability framework of the day. For example, prior to the introduction of Progress 8, GCSE grades were "scored" using a scale that ranged from 16 points for grade G to 58 points for grade A* with the intervening grades scored at 6 point intervals. From 2016, they were scored 1 point for grade G to 8 points for grade A*. This was in fact the original scoring system used until 2003. It was a brief respite since the scores changed again in 2017 to accommodate reformed GCSEs which were graded on a different scale.

In addition, the response to the Wolf Review led to changes in equivalence for some qualifications and others no longer being counted at all in school performance tables from 2014 onwards (Burgess and Thomson, 2019).

We therefore undertake a substantial amount of transformation to recalculate pupil (and therefore school) performance indicators onto a consistent basis. There are two main aspects to this. Firstly, we calculate indicators for 2014 to 2017 using the 2013 Performance tables methodology. This means we include all approved qualifications, not just those that were deemed eligible following the Wolf Review, and apply the points scoring system that prevailed in 2013 to results from 2016 and 2017. There is a necessary caveat here: schools would have responded to the prevailing accountability incentives to have entered particular qualifications. We cannot readily adjust for these different decisions, however we can attempt to use outcome measures that we believe are relatively stable in our analysis.

3.3 Defining 'Borderline' pupils

We cannot know which pupils a school thought of as being borderline. Such a judgement likely included inputs from internal low-stakes tests, teacher assessments and so on, together with decisions about how much intervention can be delivered with the resources available to a school. However, we produce a proxy, based on national Key Stage 2 tests and two relevant time-invariant factors, gender and month of birth, that predict GCSE scores well. We estimate for every pupil in our dataset their probability of achieving five or more A*-C grades including English and maths (5ACEM), and define a range of the distribution of fitted probabilities as distinguishing borderline pupils. We assume that this measure is correlated with schools' own views.

We run this in two ways, that differ principally in the assumptions made about GCSE grade inflation. First, we take an *ex post* approach and estimate the probability of 5ACEM retrospectively. Taking each year in the dataset in turn, we use logistic regression to estimate the probability, looking backwards at the relationship between these factors and actual GCSE scores in each year. Second, we take an *ex ante* approach, looking forwards and not using actual GCSE scores for each cohort. Instead, we use estimated models for an older cohort and apply those coefficients to the Key Stage 2 attainment and demographics. For example, we use coefficients from the 2010 Year 11 cohort to calculate the prevailing probability of 5ACEM at the time the 2013 Year 11 cohort were about to start Year 10. The interpretation of the difference is this: the *ex ante* approach assumes schools make no adjustment for possible grade inflation, and the *ex post* approach assumes they make full adjustment; no doubt the truth lies somewhere in between.

In fact, in this particular case, we can get a little closer to the actual information that schools had available to make their decisions. An educational organisation[vi], the Fischer Family Trust (Fischer Family Trust, 2020), provided precisely the results from the *ex ante* forward-looking estimates to schools at the time. Around 80% of secondary schools subscribed to the service during the period we analyse.

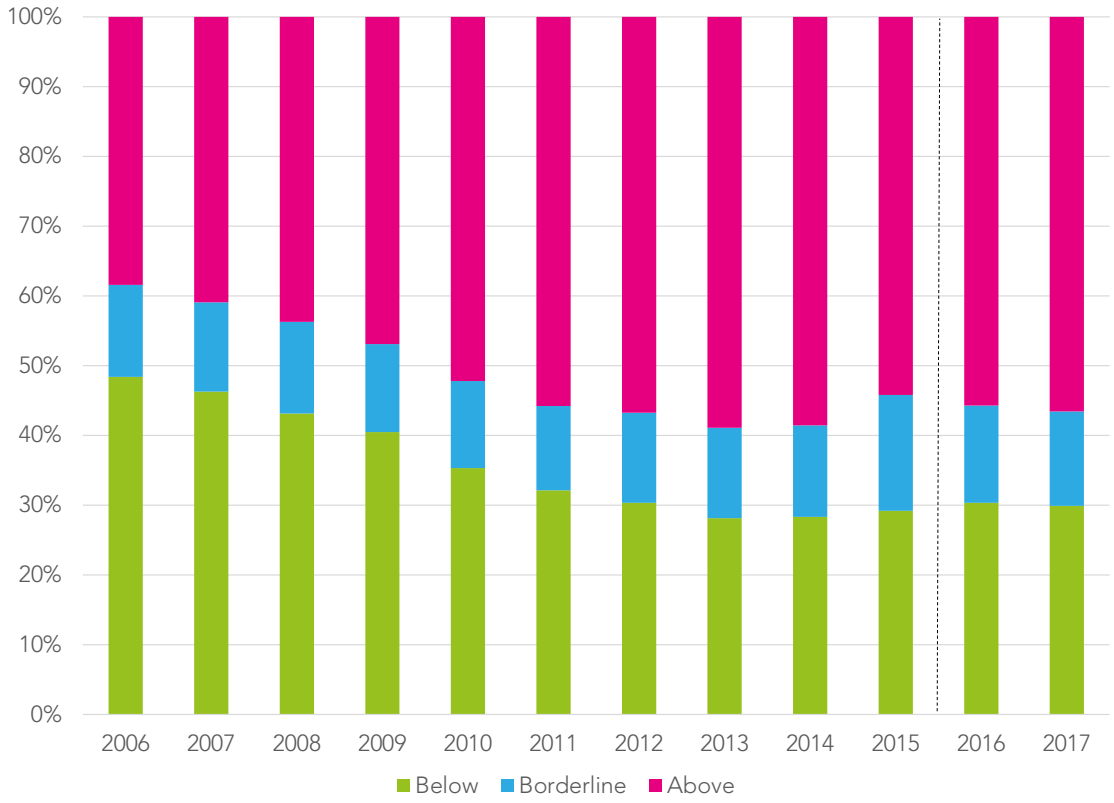
In our main specification we classify pupils as being 'borderline' if they have a probability of achieving 5ACEM between 40% and 60%. We refer to pupils with a higher probability as being in the 'above' group and those with a lower probability as the 'below' group. In our tests of robustness, we also show the effect of widening (and narrowing) this window.

Defining the borderline variable allows a few "researcher degrees of freedom" (Simmons et al, 2011) so we present below our checks on whether any of our decisions are pivotal to the results.

The outcome of this modelling is displayed in Figure 3 for the full length of available data: the percentage of pupils in each of these three groups in each year[vii] (based on *ex post* probabilities). The percentage of borderline pupils is relatively stable over the period 2005 to 2017, generally forming 12% to 15% of each cohort between 2005 and 2014 and 17%-18% of each cohort in 2015 to 2017. However, the percentage of pupils in the 'above' group increased between 2005 and 2013 as overall Key Stage 4 attainment increased (see below). This means that the group of borderline pupils is located at a relatively lower part of the Key Stage 2 distribution in the later years of the series. The borderline group was

slightly larger in size in 2015 due to the Key Stage 2 test boycott of 2010. Less granular teacher assessment data is used in place of test data (where available) which introduces more uncertainty into the calculation of *ex post* probabilities.

Figure 3: Percentage of pupils by group



If we cut the data by school, we can see that over our analysis window, 2012-2017, the variation between schools in their fractions of borderline students is not large. The 10th percentile of fraction borderline students is always 8%-9%, and the 90th percentile is 18%-19%. The 2015 cohort was affected by the 2010 Key Stage 2 boycott and so appears to be an outlier. Very few schools have hardly any borderline students and in very few schools do they account for more than a fifth (Table 1).

Table 1: Selected percentiles of school-level percentages of borderline pupils, 2012-2017

Percentile	2012	2013	2014	2015	2016	2017
10	8%	8%	8%	10%	9%	9%
25	11%	11%	11%	13%	12%	11%
50	13%	13%	14%	16%	14%	14%
75	16%	16%	16%	21%	17%	17%
90	18%	18%	18%	25%	19%	19%

3.4 Outcome variables

We examine the effects of the policy change on a number of pupil outcomes related to attainment, qualification entry and completion of Key Stage 4. We describe each in the following sections.

3.4.1 Attainment

As we outlined in Section 2, there were various changes to the secondary school accountability framework in the years preceding Progress 8. This makes it difficult to find indicators of pupil attainment that are stable over the period before and after its introduction, even after undertaking a substantial amount of recalculation (Section 3.2).

Our approach therefore is to use three indicators that have been relatively stable over this period:

- Average points score in English and maths (English and maths APS)
- The achievement of five or more A*-C grades (or equivalent) including English and maths (5ACEM)
- Mean grade in GCSEs (Mean GCSE)

English and maths APS is our primary outcome. Grades in GCSE English language (or combined English language and literature[viii]) are converted into points[ix] and then averaged. GCSEs were graded A*-G since their inception in 1988 until 2015/16. The appearance of reformed GCSEs in Key Stage 4 data for 2017 causes us a headache for our APS measure. However, we use a simple transformation to map the new 9-1 grades onto the previous points scale (see Appendix 1). The new grades were designed such that grades 3-1 correspond to the former D-G range, 6-4 corresponds to the former B-C range and 9-7 corresponds to the former A*-A range.

5ACEM was a 'high stakes' (West, 2010) indicator for schools until 2014/15, arguably the highest stakes of all indicators. Despite the introduction of Progress 8, it did not disappear entirely. The government continued to publish the percentage of pupils achieving A*-C (9-4) in GCSE English and maths[x] although as we discuss in Section 2.5, its prominence was diminished. That said, achieving A*-C in English and maths remained seen as critically important for *pupils* as it was considered essential for further study and longer-term outcomes. Since 2015, pupils have been required to retake English and maths after 16 if they do not have A*-C (now 9-4) passes.

We also use a lower stakes indicator that is less affected by the importance attached to English and maths, mean GCSE grade. Although it has been published since 2010/11, it tends to have lower prominence in comparison to other indicators. English and maths APS sits somewhere in the middle. Although not published, scores in English and maths compose 40% of the Attainment 8 measure. Furthermore, almost all pupils enter English and maths and this has been the case for the full period of our analysis. By contrast, the popularity of other subjects and qualifications has ebbed and flowed in response to changes to the accountability regime.

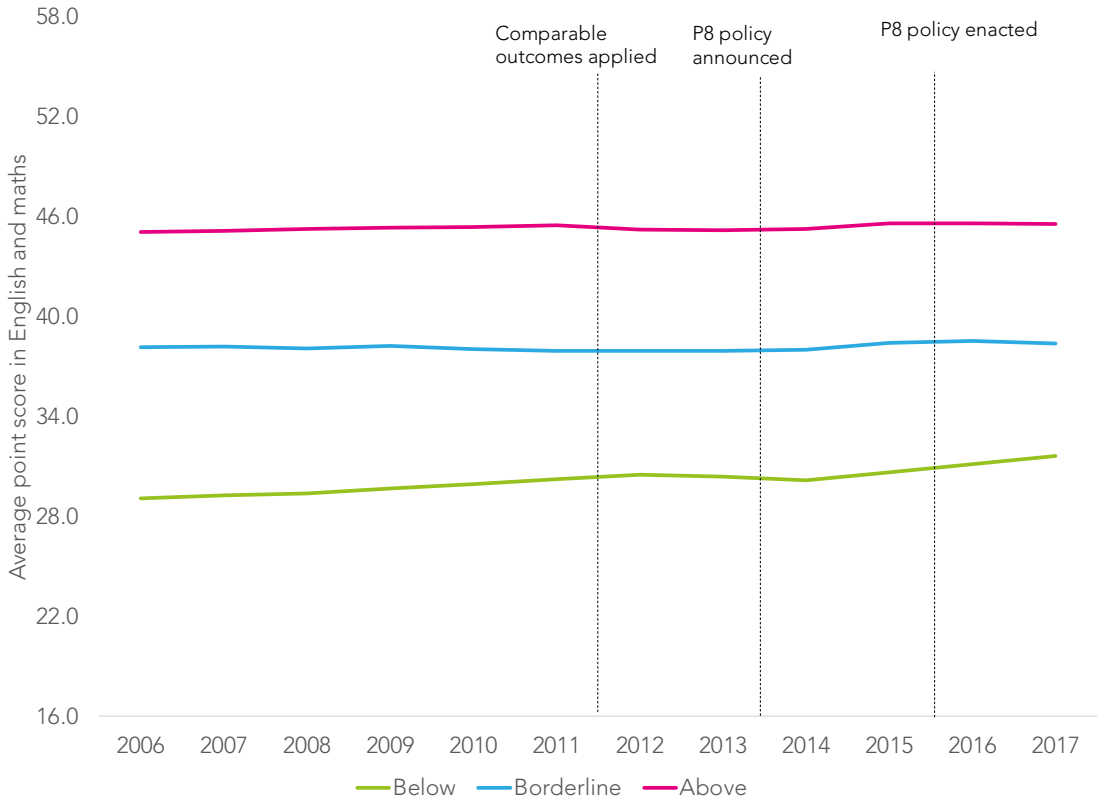
The means and standard deviations of the three key outcome measures for all pupils at the end of Key Stage 4 in state-funded mainstream schools are shown in Table 2.

Table 2: Means and standard deviations of Key Stage 4 outcomes, 2006-2017

		English and maths APS	% 5 A*-C with English and maths	Mean GCSE grade (all subjects)
2006	Mean	36.44	45%	35.83
	Std. Deviation	12.09	50%	11.52
2007	Mean	37.00	47%	36.21
	Std. Deviation	12.02	50%	11.32
2008	Mean	37.60	49%	37.01
	Std. Deviation	11.79	50%	10.94
2009	Mean	38.28	52%	37.58
	Std. Deviation	11.51	50%	10.58
2010	Mean	39.20	57%	38.18
	Std. Deviation	11.10	50%	10.23
2011	Mean	39.86	60%	38.48
	Std. Deviation	10.79	49%	10.14
2012	Mean	39.91	61%	38.68
	Std. Deviation	10.26	49%	9.89
2013	Mean	40.12	62%	38.6
	Std. Deviation	10.03	48%	9.85
2014	Mean	40.06	62%	38.68
	Std. Deviation	10.33	49%	10.04
2015	Mean	40.04	60%	38.94
	Std. Deviation	10.35	49%	10.07
2016	Mean	40.23	60%	39.04
	Std. Deviation	10.18	49%	9.99
2017	Mean	40.38	60%	39.41
	Std. Deviation	9.86	49%	10.08
All years	Mean	38.94	56%	38.01
	Std. Deviation	10.79	50%	10.45

Trends in our primary outcome, English and maths APS, for the 2006 to 2017 period for the three groups of pupils are shown in Figure 4. The attainment of the "borderline" and "above" groups is broadly stable over the whole period but there is a general increase in attainment among pupils in the "below" group.

Figure 4: Average point score in English and maths by pupil group, 2006 to 2017



3.4.2 Completion

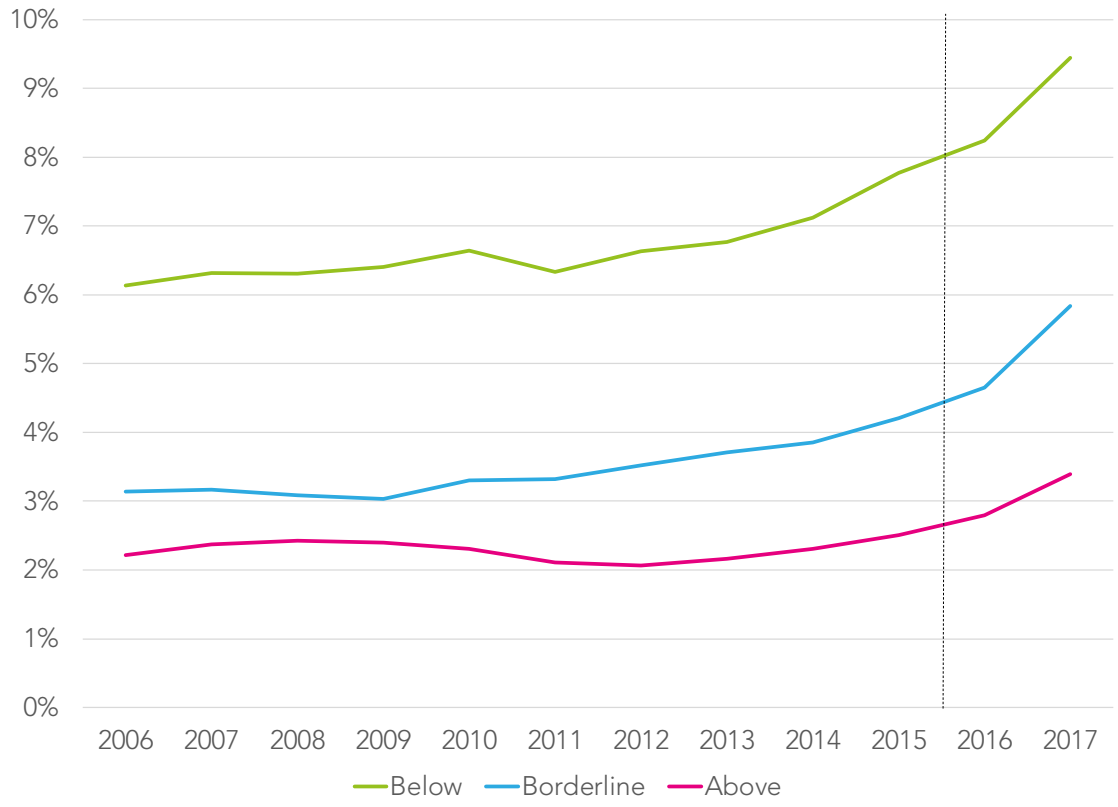
The majority of our analysis focuses on pupils who reached the end of Key Stage 4 in state-funded mainstream secondary schools. However, not all pupils who begin their secondary education in a state-funded mainstream secondary school in England complete it there. Either by the result of choice, permanent exclusion or otherwise, they may complete in another type of establishment. This includes special schools, alternative provision and independent schools. Some leave England, either to other parts of the UK or overseas.

Prior to the introduction of Progress 8, some suggested (Nye, 2017) that the new measure might shift the incentive from low-attaining pupils to pupils at risk of achieving a low value added score, particularly those who take no examinations at all. Clearly, there is some overlap between these two groups. However, pupils with low prior attainment who achieve better results at KS4 than pupils with similar prior attainment would be rewarded under Progress 8 even though they may ultimately still be low achieving relative to the national average. By contrast, pupils with high prior attainment with poor Key Stage 4 outcomes relative to similar pupils would score poorly under Progress 8, even if their outcomes are close to the national average.

We look at two measures of Key Stage 4 completion over the period 2006 to 2017. Firstly, whether pupils complete Key Stage 4 at any state-funded mainstream secondary school in England. Secondly, whether they complete Key Stage 4 at the state-funded mainstream where they were enrolled in Year 8. Our population is all pupils enrolled in Year 8 in state-funded mainstream secondary schools according to the January School Census from 2002 to 2013. These pupils would be expected to complete Key Stage 4 (Year 11) four years later, i.e. 2006 to 2017. We therefore observe their destinations four years later from both the January School Census and in Key Stage 4 attainment data. Pupils are considered to have completed Key Stage 4 in a state-funded mainstream secondary school if we find a corresponding record in either destination dataset. We do not count transfers to university technical colleges (UTCs) and studio schools, new types of state-funded mainstream school introduced from around 2012 onwards, which cover the 14-19 range because these schools would not have been available to Year 8 pupils.

Summary completion data for 2006 to 2017 is shown in Figure 5.

Figure 5: Percentage of pupils enrolled in state-funded mainstream schools in Year 8 who do not complete Key Stage 4 within the sector

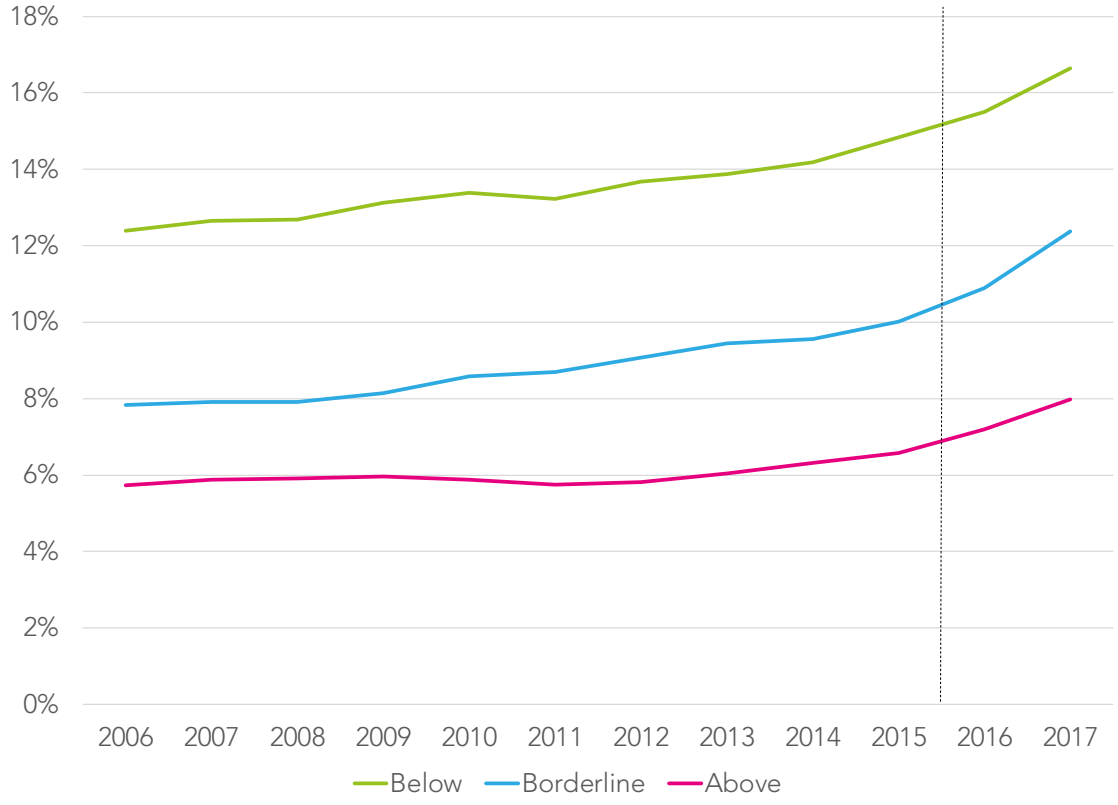


Three features stand out from Figure 5. Firstly, it has always been the case that a percentage of Year 8 pupils do not complete Key Stage 4 in the state-funded sector. Reasons for this include emigration and moving to special schools, independent schools or alternative provision. This has particularly been the case for pupils in the "below" group. Secondly, rates of non-completion were increasing prior to the introduction of Progress 8.

Finally, it appears that rates of non-completion increased more rapidly following its introduction, particularly between 2016 and 2017. However, this may be the result of increased places in educational establishments that admit at age 14 such as UTCs and studio schools.

We also consider the percentage of pupils who change school between Year 8 and the end of Key Stage 4. This includes pupils who move from one state-funded mainstream to another as well as pupils included in Figure 5. We would expect some pupils to move schools in any event. In some parts of the country, a three-tier system operates in which pupils transfer schools twice, typically at the end of Years 3 or 4 and then again at the end of Years 7 or 8. We therefore do not include pupils moving through a three-tier system when we analyse the percentage of pupils who complete Key Stage 4 at the same school where they began Year 8. Summary data for remaining pupils is shown in Figure 6.

Figure 6: Percentage of Year 8 pupils attending state-funded mainstream schools who change school before the end of Key Stage 4



Patterns of school changes after Year 8 (Figure 6) are broadly similar to patterns of exits from the state-funded mainstream sector (Figure 5). Pupils in the "below" group are the most likely to be affected, rates were increasing prior to the introduction of Progress 8, and rates appear to have increased by a greater degree subsequently.

3.4.3 Qualifications Entered

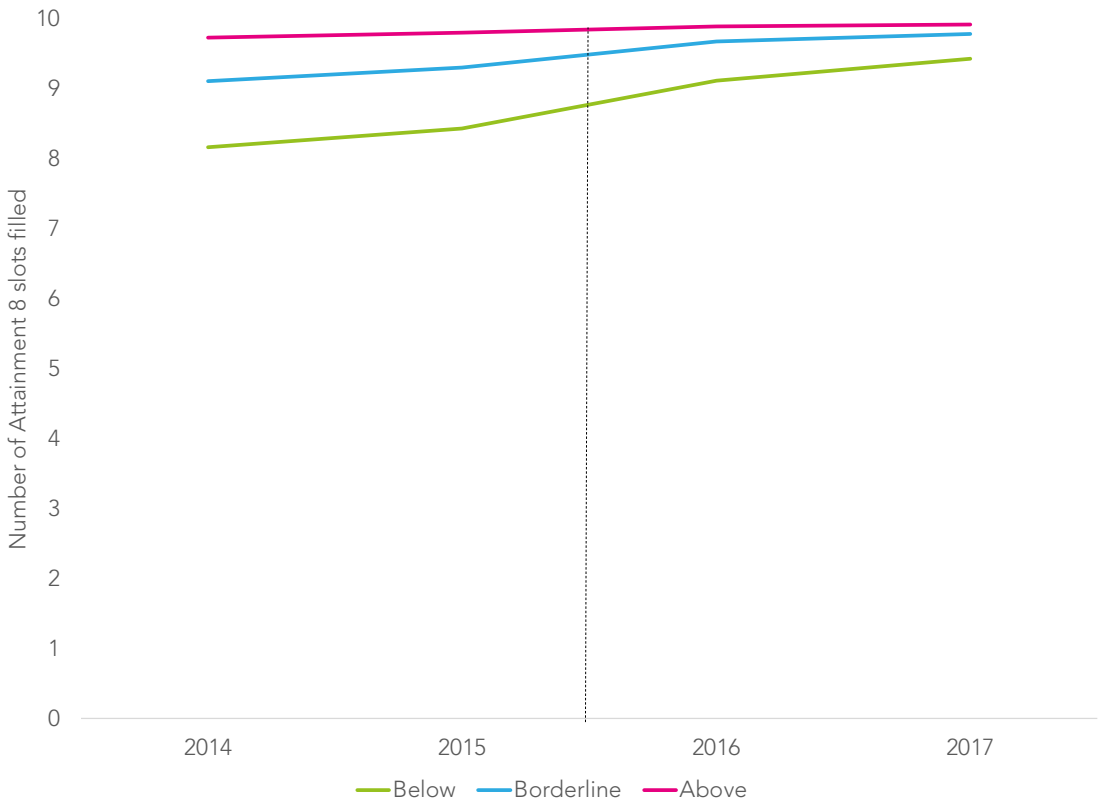
We also examine the effect of the introduction of Progress 8 on the qualifications entered by pupils.

As Figure 2 showed, the number and nature of qualifications entered has tended to change over time. The Wolf Review of 2011 led to particularly large changes when implemented for the 2014 Performance tables.

We calculate the number of Attainment 8 "slots" filled by pupils as a measure of qualifications entered. This has a maximum value of 10: two for English, two for maths, three for other subjects counted in the English Baccalaureate (science, humanities, and languages) and then any three other subjects. Attainment 8 was first introduced in 2016 but we retrospectively calculate the measure for 2014 and 2015. We do not calculate it for earlier years as we would have to make subjective decisions about which qualifications would have been counted had the Wolf reforms been implemented earlier than 2014.

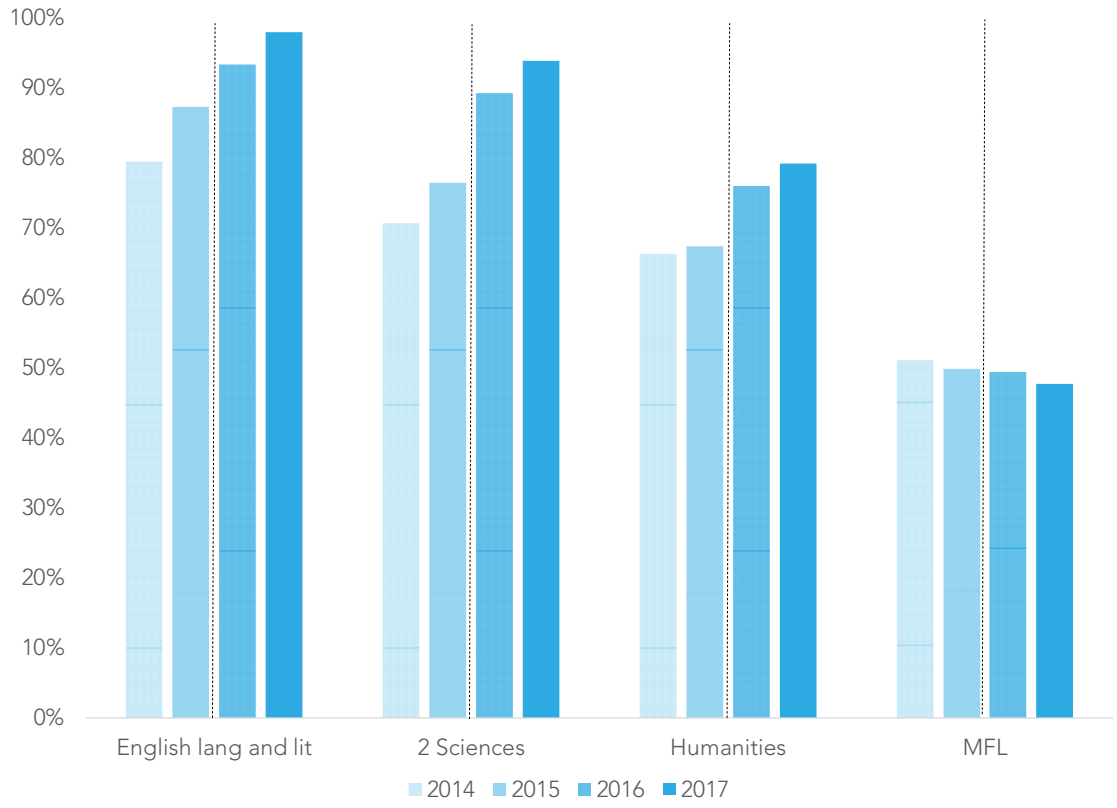
Trends in entry rates are shown in Figure 7. There is only a slight uplift in the entry rate of the "above" group: 85% of this group entered the maximum of 10 slots in 2014. Entries increased by 0.5 slots on average among pupils in the borderline group and by 1 slot on average among pupils in the "below" group between 2015 and 2017.

Figure 7: Number of Attainment 8 slots filled by pupil group, 2014 to 2017



The increase in slots filled was largely driven by increased take-up of the Ebacc subjects. There was an incentive under Attainment 8 to enter pupils in GCSE English literature as well as English language as the best grade would be doubled if both were entered[xi] and to fill all three slots assigned to the other Ebacc subjects. The effect of this is shown in Figure 8. The percentage of pupils entering two or more GCSEs in science[xii] and the percentage entering a GCSE in a humanity[iii] were increasing prior to the introduction of Progress 8 but increased by a greater margin thereafter. This was not the case in modern foreign languages (MFL), however. Progress 8 did nothing to halt its decline.

Figure 8: Percentage of pupils entered in GCSEs in selected subjects, 2014 to 2017

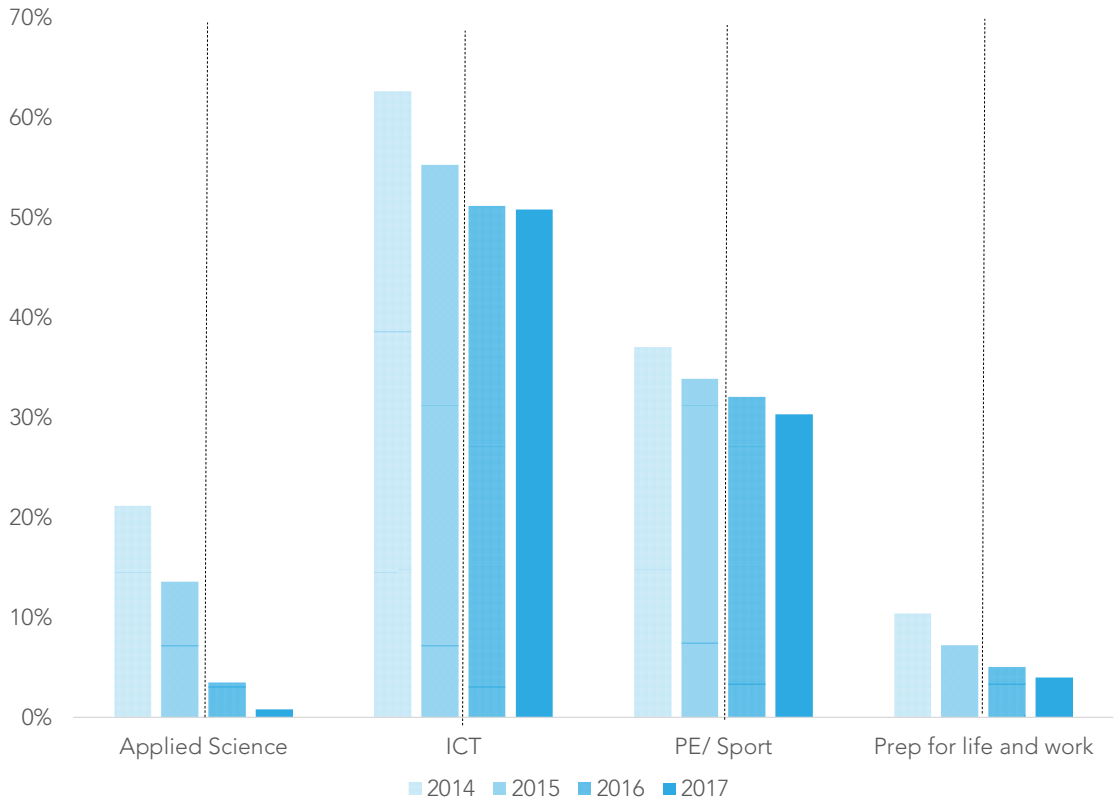


It was not the case that pupils simply started taking more qualifications, rather they changed the sorts of qualifications they were entering. In Figure 9 we show entry rates fell over the same period in four subject areas: applied science, information and communications technology (ICT), physical education (PE) and sport, and preparation for life and work. These charts include GCSE and other equivalent qualifications. Entries were decreasing immediately prior to the introduction of Progress 8 and continued to fall thereafter.

Applied science includes the hitherto popular BTEC science course which, unlike GCSE science, did not contribute to the Ebacc slots of Attainment 8. Schools therefore tended to switch back to GCSE. Preparation for life and work includes a range of qualifications that tended to be aimed at lower-attaining pupils and which taught life skills and self development. This included the ASDAN certificate of personal effectiveness (CoPE), which

has been shown to have had some effect on attainment in other subjects (Harrison et al, 2015).

Figure 9: Percentage of pupils entered for qualifications in selected subjects, 2014 to 2017



3.5 Analysis period

The key threshold measure of five or more A*-C grades including GCSE English and maths was first published in 2005/06. Data for the period 2005/06 to 2016/17 was available at the outset of the project. For the reasons we set out here, we restrict our analysis to the period 2011/12 to 2016/17. We also test the robustness of our results to different choices in Section 4.4.

We use data from 2011/12 for three reasons. Firstly, this was the first year that comparable outcomes was officially applied to English and maths GCSEs (Ofqual, 2011). Secondly, and relatedly, the percentage of pupils in the below-borderline group largely levelled out as indicated in Figure 3. Thirdly, there appear to be roughly parallel trends between the three groups during this period as shown in Figure 4. The difference-in-difference methodology we adopt to estimate the causal effect of Progress 8 on attainment, described in the following section, rests on stable common trends between the three groups prior to its introduction. We examine the latter in more detail in Section 4.2.

Having determined the analysis period, we standardise the outcome variables related to attainment (section 3.4.1) to standard deviation units over the analysis period 2012 to 2017. This means our results are presented in effect size units.

3.6 Identification of a causal effect

Our difference-in-difference approach works off a policy change interacted with pupils in different groups, likely to be differentially affected by the policy. These groups are pre-determined for the reform: characterized by pre-reform definitions and estimated using only pupil characteristics that were fixed before the reform. The key group are those considered to be borderline for achieving the key accountability metric pre-reform, pupils who were marginal to the sharp threshold of five A*- C grades.

We estimate the following as our main specification for pupil *i* in school *s* in academic-year *t*:

$$g_{ist} = \beta.X_i + \mu_{st} + \delta.\tau + (\sigma_0 + \sigma_1.\tau) * b_{is} + (\alpha_0 + \alpha_1.\tau) * a_{is} \quad (1)$$

The dependent variable, g_{ist} is the test score outcome for a pupil (English APS, 5ACEM or Mean GCSE grade). The standard diff-in-diff terms are group dummies b_{is} and a_{is}, a_{is} is equal to 1 if pupil *i* in school *s* is denoted "above the borderline"; and b_{is} is equal to 1 if pupil *i* in school *s* is denoted "below the borderline", and τ as the "after" dummy with τ =1 in the post-reform years (2016 and 2017), and zero otherwise. The key difference-in-difference terms are $\tau.b_{is}$ and $\tau.a_{is}$. We are able to supplement these in our data with pupil characteristics, and school-year fixed effects. These are: X_i, a vector of pupil covariates[xiii] (Key Stage 2 attainment, gender, ethnicity, free school meal eligibility in GCSE year, month of birth, first language and interactions between them), and μ_{st} is a set of fixed effects representing each school s in each year t. We standardise Key Stage 2 scores for each year and fit them as a third-degree polynomials (cubic).

The difference-in-difference parameters are σ_1 and α_1: these indicate the differential impact of the reform on test score outcomes.

The identification of causal effects using the difference-in-difference model set out in equation 1 assumes parallel trends in outcomes between the groups of pupils prior to the policy change. This is one reason our analysis period starts in 2011/12 (section 3.5). We examine the pre-reform trends in greater detail in Section 4.2. The model can be extended to allow for group-specific trends (see for example Angrist and Pischke, 2014) in our multi-year context:

$$g_{ist} = \beta.X_i + \mu_s + \delta.\tau + (\sigma_0 + \sigma_1.\tau) * b_{is} + (\alpha_0 + \alpha_1.\tau) * a_{is} + \gamma_j.y_{jt} + \theta.a_{is} * t + \varphi.b_{is} * t \quad (2)$$

or

$$g_{ist} = \beta.X_i + \mu_s + \delta.\tau + (\sigma_0 + \sigma_1.\tau + \varphi.* t) * b_{is} + (\alpha_0 + \alpha_1.\tau + \theta * t) * a_{is} + \gamma_j.y_{jt} \quad (2b)$$

Here we have added group specific time-trends, $a_{is} * t$ and $b_{is} * t$, and a set of common year effects, $\gamma_j.y_{jt}$.

3.7 School survey

In order to understand a little more about how school behaviour changed in response to Progress 8, we invited responses from schools to an online survey. We are cautious about reading too much into the results given that we were asking respondents to recall events from four to five years earlier. However, they do reveal something about teachers' post-hoc perceptions of the impact of Progress 8 on the organization of teaching and learning.

3.7.1 Administering the Survey

The survey covered four broad themes: curriculum organisation, timetabling decisions, the allocation of teachers to classes (where subjects use sets) and the provision of intervention sessions. Respondents were asked about the current situation in their school. Those who were at the same school in the 2014/2015 academic year, the year before Progress 8 was introduced, were asked a series of questions about how things had since changed. These changes may not have been due to Progress 8, of course. Other policy changes, such as the introduction of 9-1 GCSEs, and reducing per-pupil funding may also have played a part (Sibieta et al, 2019).

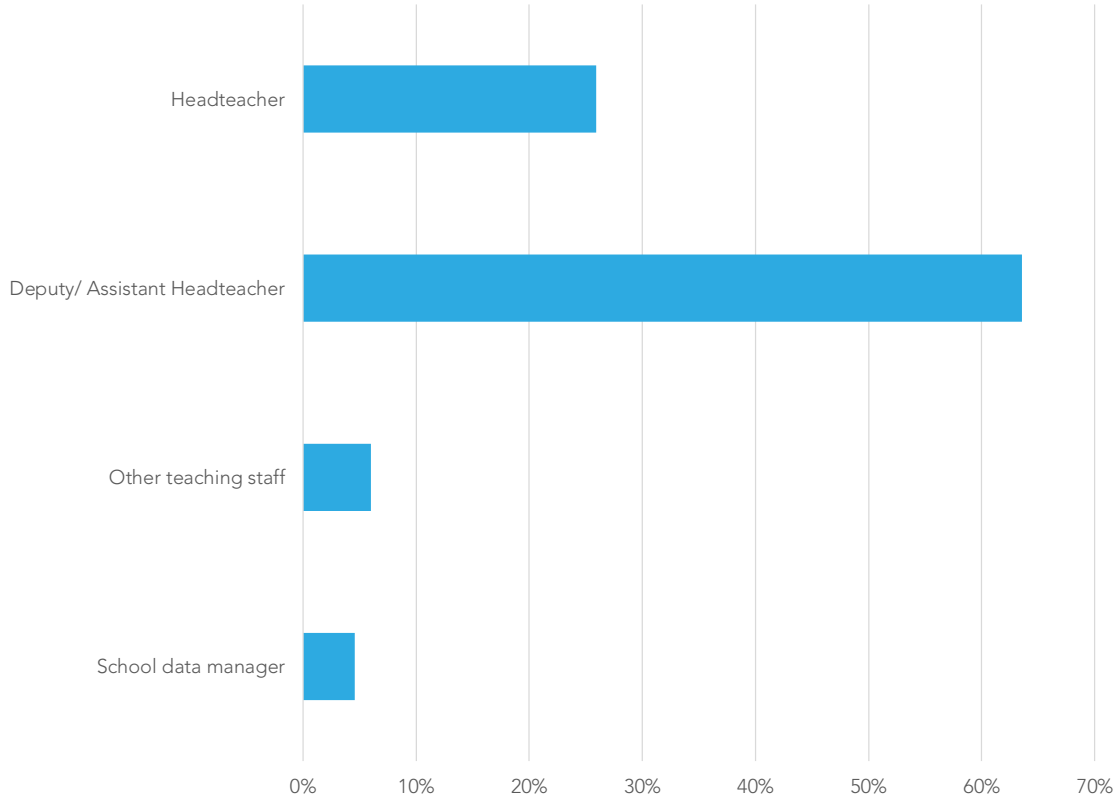
We administered the survey in Autumn 2019 using the Survey Monkey platform. A sampling frame of 2,875 schools was established. These all had end of Key Stage 4 results for at least 10 pupils in both 2015 and 2017 and were still open in October 2019. As we wanted to include survey response data in our main analysis, we tracked responses using a school identifier. Surveys were sent to schools using a list of email addresses of subscribers to the FFT Aspire system who had given consent to be contacted. We sent surveys to one user in each school, preferring responses from the earliest subscriber[xiv]. In total, we sent surveys to 2,054 schools.

There is a risk with this approach that the achieved sample of schools will be more likely to respond to accountability incentives. However, we show in Section 5.6 that overall results for the sample do not differ substantively from the overall for all schools presented in Section 4.1

3.7.2 Response Rates

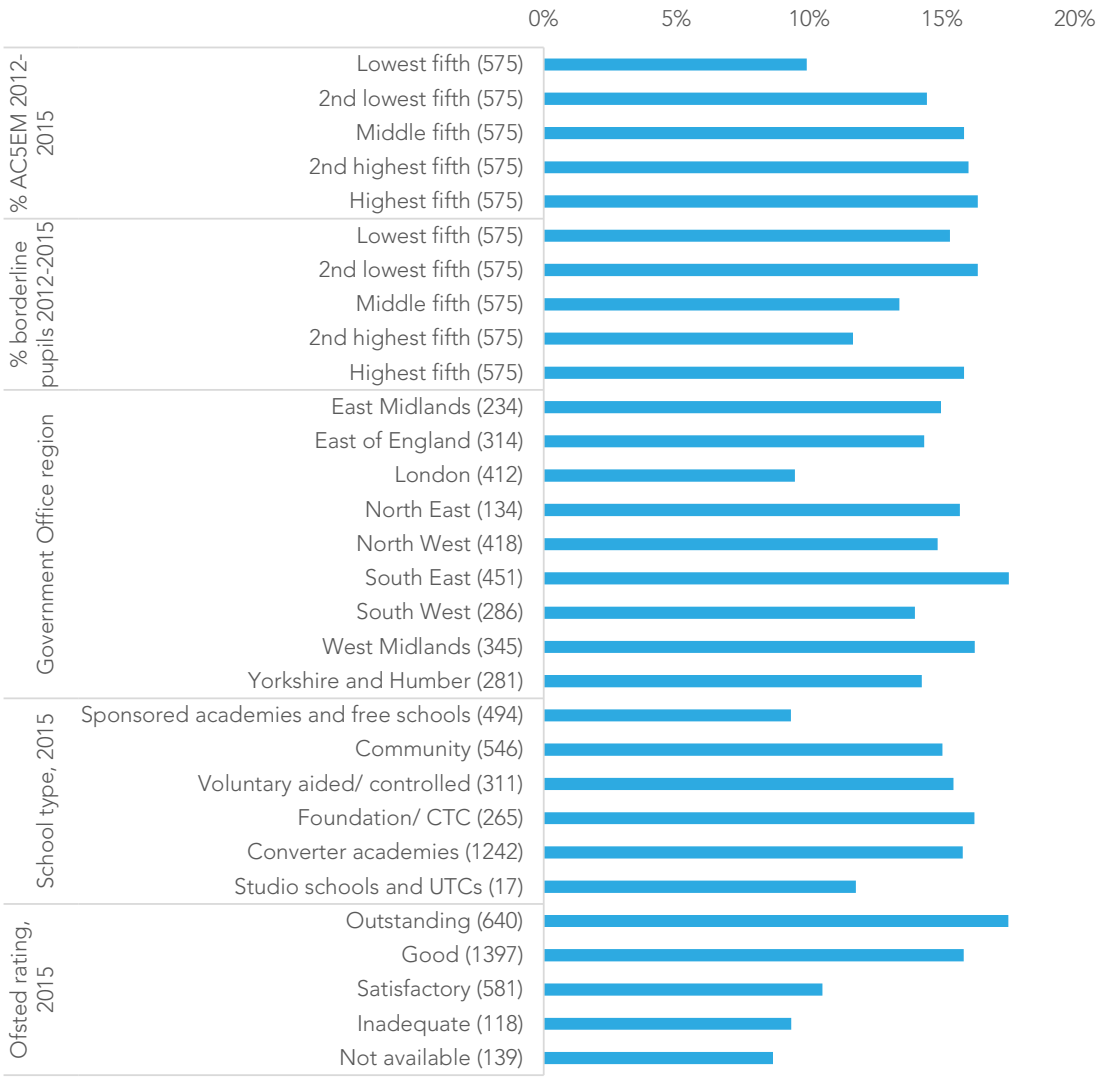
274 schools responded to the initial survey. We then surveyed replacement teachers amongst non-responding schools. This produced a further 143 responses. In total, 417 schools responded. This constitutes a response rate of 14.5% of the full sampling frame. Of these, 298 (71%) were working at the same school in the 2014/15 academic year. As Figure 10 shows, almost 90% of respondents were members of the school leadership team.

Figure 10: Current role of survey respondents



Some types of school were less likely to respond than others. These include sponsored academies, schools rated less than good at the start of 2015, those in London and those in the lowest fifth for attainment in 5ACEM during the pre-reform period (2012 to 2015). This is summarised in Figure 11.

Figure 11: Response rates by school characteristics



4 Results

We use six years' data from 2011/12 through to 2016/17 on all the pupils in England with KS4 outcomes. For our main analysis, we have a final sample of 3.1 million pupils in 3,165 schools.

There are three reasons why we choose to start the analysis period in 2011/12. Firstly, this was the first year that comparable outcomes was officially applied to English and maths GCSEs (Ofqual, 2012). Secondly, together with the levelling out of the trend for the below-borderline group in Figure 4**Error! Reference source not found.**, leads us to use data from years 2012 to 2015 as the pre-reform years in our analysis.

First of all, we describe what we might have expected to see if the policy change were indeed to have an effect on GCSE scores. We then test the assumptions necessary to interpret the difference-in-difference estimates we subsequently estimate as causal impacts of the policy that introduced Progress 8.

We then present the results of our primary outcome, the average point score in English and maths converted to standard deviation units. Second, we test the robustness of those results to alternative specifications. Third, we consider alternative outcome measures. Fourth, we explore heterogeneous responses to the reform from schools in different circumstances and schools of different types. Finally, we supplement this large-scale quantitative analysis with some qualitative work summarising the responses of schools to a questionnaire.

4.1 What results could be expected?

Before the policy change was announced, there were long standing incentives for schools to favour the borderline group, as discussed above. We would therefore expect to see a change in attainment for other pupils relative to the borderline group after the reform. However, as we outline in Section 3.4, this disjuncture may not be as ideal as we would wish as there remained strong incentives in place for pupils to achieve grade C (later grade 4) passes in English and mathematics.

The transition period began in October 2013 with the announcement of the policy, with the new performance measure to be first applied for outcomes in the year 2015/16. How might schools react? By October 2013, it seems likely that at least some of the big prioritisation decisions for the year 2013/14, not least the allocation of teachers to classes, would have been taken, and so we would not really expect any impact on attainment for the 2013/14 Year 11 cohort. It seems more likely that schools would be able to change policies (if they wished to) from 2014/15. This might only affect the schools most attuned to the incentive structure and keenest to change. This would proceed as follows: the performance measure policy change could affect the exam outcomes:

- in 2014/15 after one year of change in schools' prioritisation decisions, presumably for year 11 students;
- in 2015/16 after two years of change in schools' prioritisation decisions, presumably for years 10 and 11 students;
- in 2016/17 after three years of change in schools' prioritisation decisions, presumably for years 9, 10 and 11 students.

The period we could describe as fully post-policy-change would be from when all secondary school years were under the new regime, which would start with the 2019/20 GCSEs. From then, there would be a new prioritisation regime, and a new 'steady state' of school plans; at least, there would in principle, but further important reforms, such as the introduction of reformed GCSEs, have continued to arrive and potentially affect schools' plans. The implication of the new performance regime is that for schools, incentives to invest in pupils of different abilities are now 'equal' across all pupils.

What are the implications of this for the results? Under the hypothesis studied here that schools react in an optimising way to the accountability framework they work under, we

would expect to see a gradual build-up of change as schools switch to the new investment strategy and pupils have more and more years under the new approach. Note that such a gradual change is simply due to the passage of time, rather than any slow or reluctant reaction by schools; it takes two years for pupils to have two years of priority investment. We expect to see zero change in exam outcomes in 2013/14 exams, through a small effect in 14/15, bigger in 15/16 and bigger still in 16/17 (and bigger yet in 17/18), until outcomes stabilise.

Note that this is a very different expected profile to the archetypical difference-in-difference model in which the policy change produces an instant and on-going effect (see for example the many figures illustrating this in Angrist and Pischke (2015, figures in chapter 5) and Cunningham (2018, figures in chapter 10)).

This has important implications for our evaluation of the fit of our model, principally in terms of the standard analysis of prior trends and placebo tests. Essentially, the issue is this: the optimising model of schools reacting to new incentives implies (as above) a necessarily gradual reaction to the policy, some portion of which will necessarily happen shortly before the actual implementation. In this case, the data will present as differential prior trends and pre-implementation effects. But these patterns are precisely those that are taken to cast doubt on the validity of a difference-in-difference analysis. One solution would be to count the policy change date as the announcement of the policy, October 2013, but this runs up against the data problem that that moment is only one year after the adoption of "comparable outcomes" policy that ended grade inflation and instituted cohort referenced marking of GCSEs, potentially prejudicing the before/after comparison.

Because of these issues, we first take a cautious approach and present graphs showing a relatively long time series of effects on GCSEs for the groups of pupils we have defined above. We discuss their interpretation. We then present the more formal and standard difference in difference analysis, which can be interpreted with the time series of effects in mind. We also present a set of robustness analyses for those results.

4.2 Graphical analysis of GCSE impacts over time

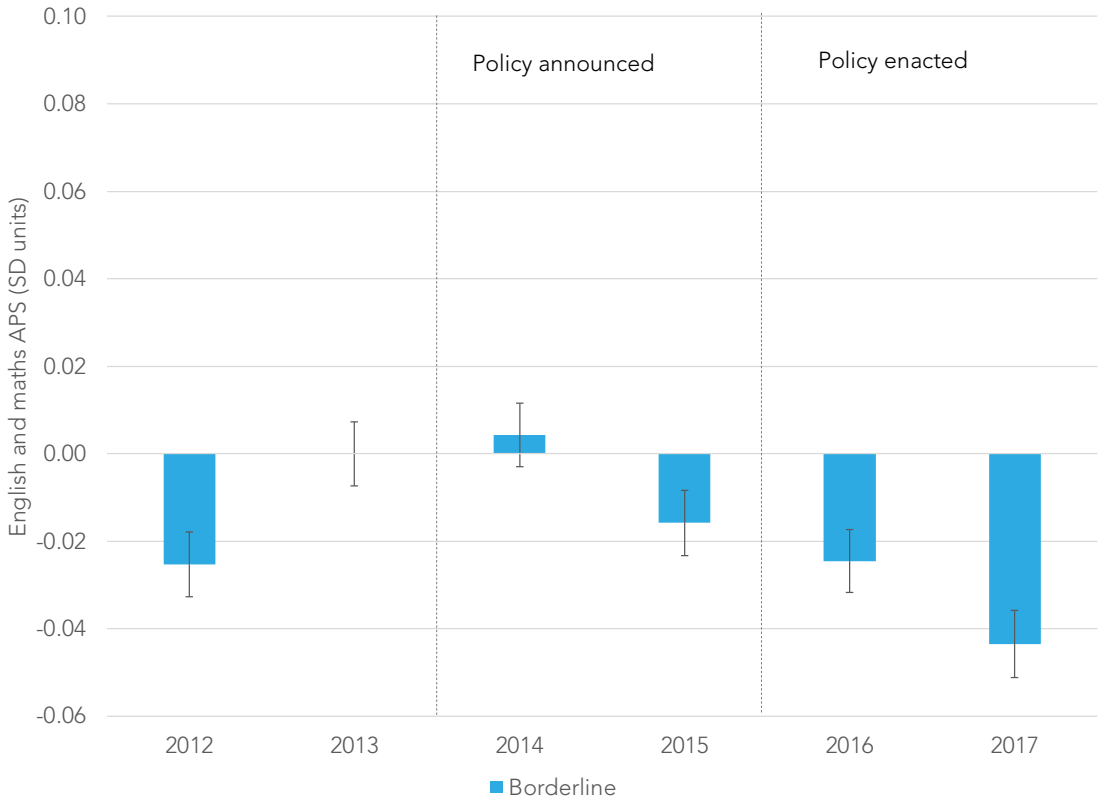
The two key assumptions for a difference-in-difference approach to yield a valid causal estimate are that there is no movement between groups, and that the different groups considered have common outcomes trends before the policy change. By definition, there can be no movement of a pupil between groups after the reform, as that derives from our estimation. We now address the issue of prior trends.

Figure 12 shows the results of estimating:

$$g_{ist} = \beta . X_i + \mu_s + \lambda_t + \sum_{\tau=2012}^{2017} \delta_\tau . B_{i\tau} \quad (3)$$

where we present the coefficient (δ) for the borderline group (B) interacted with each year in turn 2011/12 through 2016/17 (with 2012/13 acting as base year), along with the associated 95% confidence intervals clustered by school and year.

Figure 12: Difference-in-difference estimates for the borderline group by year



The pattern from 2014/15 onwards fits with the hypothesis set out above, a gradual decline each year after 2013/14 for the borderline group. This is consistently downward, but particularly marked in 2016/17. It is also worth noting that there is some instability in our outcome measure prior to 2014/15 as a result of the rise and fall in multiple entry (resits) in English and maths (Appendix B). The result for 2017 (-0.04 SD) is equivalent to around 6% of a grade at GCSE.

Figure 12 compares the borderline group of students with the single category of all other students. This thereby restricts the impact of the policy on all other students to be the same. It seems possible that this would not be the case, so we now separate out the above-borderline group and the below-borderline group as we specified above. This more flexible approach is shown in Figure 13. To recap, the mirror image of the borderline group declining in response to the policy is that at least one of these groups should increase as they are no longer incentivised against.

Figure 13: Difference-in-difference estimates for the above and below groups by year

We see that indeed both groups tend to gain from 2013/4, the above group marginally, the below group more dramatically so. This is consistent with the hypothesis that changing prioritisation policies of schools will have more effect each year from 2013/4 onwards as pupils are "treated" for successively more of their school careers.

To summarise, these two figures show patterns that are consistent with the hypothesis that the reform to the accountability system changed schools' incentives for targeting interventions and that this in turn led to changes in pupil outcomes. These changes worked against the 'borderline' group, and mildly in favour of the 'above borderline' group and strongly in favour of the 'below borderline' group. We are very clear, however, that the patterns could also fit other non-causal stories, that there simply are unexplained trends starting roughly around the time of the reform we are focussing on and which are just unluckily coincident. The pattern does not look like a classic difference-in-difference graph. As has been noted (Cunningham 2018; Pischke, 2005), distinguishing between unexplained trends and a gradual causal effect with anticipation can be difficult. Working in favour of the hypothesis we set out is that the borderline group we define is quite narrow and "specific" – that is, it is well defined only in relation to the accountability process for schools, but is meaningless in relation to any other drivers for school behaviour.

4.3 Difference-in-difference results

Table 3 presents our difference-in-difference results. This uses average points score in English and Maths as the dependent variable and our base definition of 'borderline' as pupils with a 40-60% probability of achieving the threshold, and includes different specification of controls. Standard errors are clustered at school-by-year level. For each specification, we report coefficients on the below group, the above group, and the post-reform dummy. Note that the simple group dummies cannot be straightforwardly interpreted: they are strongly correlated with the pupil characteristics also included in the regression and are only separately identified by functional form (the logistic regression determining the borderline group). However, the main focus of interest are the difference-in-difference coefficients shown in bold in Table 3.

Column (1) reports the base model with no additional controls. This shows a post-reform increase in test scores relative to borderline pupils of 0.06 of a standard deviation (SD) for the below-borderline group, and essentially no effect for the 'above' group.

The second column adds school-by-year dummies, thereby controlling for aggregate time effects, time-invariant school effects, and school-year specific effects in a very flexible way. The difference-in-difference estimates barely change. The third column adds pupil characteristics, listed below the table, but removes the school-by-year dummies. This has two effects. First, as expected, this makes a big difference to the simple estimated group effects, as they are simply non-linear functions of some of the characteristics. Second, and more importantly, the difference-in-difference coefficient for the 'above' group now become positive and statistically significant; the coefficient for the 'below' group declines slightly.

The fourth and final column presents our full specification with both pupil characteristics and school-year effects, we find that the post-reform effect for the 'above' group is 0.01 SD and for the 'below' group is 0.057 SD. The average of these two terms weighted by the number of pupils in the above and below groups, 0.027SD, is the additional value that the borderline group experienced prior to the introduction of Progress 8.

We postpone a full discussion of these results to late in the report, but two immediate conclusions are that: (i) the use of the threshold measure made a statistically significant difference to school outcomes, we assume arising from changed school behaviour, focusing their resources on the incentivised group of pupils, and (ii) when that incentive was eliminated, schools reacted, and redistributed resources to the non-borderline groups more heavily weighted towards the below-borderline group.

Table 3: Key Parameter Estimates from Headline Models

	1		2		3		4	
	b	se(b)	b	se(b)	b	se(b)	b	se(b)
above	0.764**	0.003	0.686**	0.002	-0.032**	0.002	-0.033**	0.002
below	-0.763**	0.002	-0.744**	0.002	-0.011**	0.002	-0.016**	0.002
reform	0.035**	0.004			-0.007	0.004		
Interaction of above and reform	-0.006	0.004	-0.008*	0.003	0.009*	0.003	0.010**	0.003
Interaction of below and reform	0.064**	0.004	0.063**	0.003	0.054**	0.003	0.057**	0.003
Number of pupils (thousands)	3096		3096		3094		3094	
Number of schools	3165		3165		3165		3165	
R-squared	0.45		0.51		0.64		0.67	
pupil covariates	No		No		Yes		Yes	
school*year fixed effects	No		Yes		No		Yes	

* significant at the 5% level, ** significant at the 0.1% level

Notes

1. The outcome measure is points score in English and maths converted in standard deviation units
2. Standard errors are clustered by school*year
3. Pupil covariates are standardized Key Stage 2 score, free school meal eligibility, ethnicity, gender, month of birth, IDACI decile, first language (English/ other) and interactions of the characteristics with standardized key stage 2 score. Standardised Key Stage 2 score is fitted as a third-degree polynomial (cubic).

Table 4: Robustness checks

	1. Narrower Panel		2. Wider panel		3. Separate school/ year dummies		4. Ex-ante borderline definition		5. Wider definition of borderline (30-70%)	
	b	se(b)	b	se(b)	b	se(b)	b	se(b)	b	se(b)
above	-0.032**	0.002	-0.028**	0.001	-0.031**	0.002	-0.021**	0.002	-0.012**	0.002
below	-0.017**	0.002	-0.020**	0.002	-0.015**	0.002	-0.036**	0.002	-0.052**	0.002
Interaction of above and reform	**0.017****	**0.003**	**0.000**	**0.002**	**0.008***	**0.003**	**0.007***	**0.002**	**0.006***	**0.002**
Interaction of below and reform	**0.058****	**0.003**	**0.060****	**0.003**	**0.055****	**0.003**	**0.059****	**0.003**	**0.064****	**0.003**
Number of pupils (thousands)		2567		3626		3094		3094		3094
Number of schools		3159		3175		3165		3165		3165
R-squared		0.67		0.68		0.66		0.67		0.67

	6. Narrower definition of borderline (45-55%)		7. Excluding early adopters		8. Within-between		9. Based on legacy pupils		10. With trend parameters	
	b	se(b)	b	se(b)	b	se(b)	b	se(b)	b	se(b)
above	-0.032**	0.002	-0.034**	0.002	-0.021**	0.001	-0.041**	0.002	-0.021**	0.004
below	-0.010**	0.002	-0.013**	0.002	-0.029**	0.002	-0.013**	0.002	-0.036**	0.005
Interaction of above and reform	**0.010***	**0.003**	**0.012****	**0.003**	**0.008****	**0.002**	**0.015****	**0.003**	**0.020***	**0.006**
Interaction of below and reform	**0.050****	**0.003**	**0.052****	**0.004**	**0.058****	**0.002**	**0.057****	**0.003**	**0.016***	**0.006**
Interaction of group means of above and					**-0.006**	**0.008**				
Interaction of group means of below and					**-0.010**	**0.014**				
Number of pupils (thousands)		3094		2776		3094		3145		3094
Number of schools		3165		2839		3165		3196		3165
R-squared		0.67		0.63				0.63		0.66
Variance partition coefficient						0.08				

* significant at the 5% level, ** significant at the 0.1% level. See notes for Table 3

4.4 Tests of robustness

We made a number of decisions on data and modelling, underlying the results in Table 3. Some of these choices were driven by policy, such as the use of comparable outcomes in the awarding process for GCSE English and maths from 2012 onwards stabilising (to some extent) examination grades. This led us to start our sample from 2012. However, other choices were less clear cut, and we now check to see if any of these were overly consequential to the results by presenting in Table 4 estimates based on alternative specifications.

The two key assumptions for a difference-in-difference approach to yield a valid causal estimate are that there is no movement between groups, and that the different groups considered have common outcomes trends before the policy change. By definition, there can be no movement of a pupil between groups after the reform, as that derives from our estimation. We addressed the issue of prior trends above.

Columns (1) and (2) show the effect of increasing and decreasing the number of pre-reform years. In column (1) we add data for 2010/11, the year prior to the application of comparable outcomes in GCSE English and maths. The difference-in-difference estimates barely change for the below group but there is some slight change for the above group arising from the slight instability in the outcome measure for this group prior to 2012/13 as indicated in Figure 13.

Column (3) uses fixed effects for schools, and fixed effects for years rather than school-by-year effects above. This is based on consistent identifiers for schools (for example, linking schools that closed as community schools and re-opened as Academies). This makes little difference. Columns (4), (5) and (6) adjust the definition of borderline pupils: respectively switching to the *ex ante* definition of borderline, using a broader definition of the 'borderline' group (pupils with a chance of hitting the threshold between 30% and 70%), and a narrower one (45% and 55%). The first of these makes very little material difference. Widening the borderline group increases the effect of the reform by 0.007 SD for the below group and narrowing it reduces it by an equivalent amount. To reiterate, we can only estimate which group of pupils the school thought of as borderline, we do not know for sure.

Column (7) excludes the 327 schools which opted early into Progress 8 in 2014/15, which reduces the effect for the below group by 0.005SD. Column (8) instead adopts a multilevel modelling approach and fits the school-year effects as level 2 random effects; this allows us to model the variation *between* school-years, notably in school-year percentages of pupils in the above and below groups. The estimates for the group-level means of the percentages of pupils in the above and below groups interacted with reform shown in column are non-significant. This suggests that the effects of the reform on above-borderline and below-borderline pupils does not vary between schools with respect to the fraction of pupils in each of these groups.

In column (9) we define a sample based on pupils observed on roll in Year 8 in state-funded mainstream schools[xv] (as opposed to the criterion above based on those who complete their secondary education in such a school). Again, the difference-in-difference estimates are very similar to our main results from column (4) in Table 3. Finally, in column (10) we

show the effect of fitting a linear trend in outcomes for the above and below and group (equation 2 in Section 3.6). This increases the effect for the above group by 0.01SD and reduces the effect of the below group by 0.03 SD. This latter change could be the result of incorrectly correcting for an anticipatory effect in 2014/15 (see Figure 13).

4.5 Other outcomes

We examine the impact of the reform on a number of other outcomes and present the results in Table 5.

4.3.1 Other attainment outcomes

First, we consider the headline pre-reform indicator, whether a pupil achieves five or more A*-C grades at GCSE (or equivalent) including English and maths (5ACEM). We fit the achievement of 5ACEM as a linear probability model. We observe a post-reform effect for the 'below' group relative to the borderline group, of 0.04 of a SD. This suggests that schools' resource investment 'works', that is, it affected the headline accountability figure that schools were aiming to influence. We also observe a slight negative effect of -0.01 SD for the above group.

Second, we examine the impact on a 'lower-stakes' indicator, the mean grade achieved in GCSEs (excluding equivalent qualifications). Again, we observe a post-reform effect for the 'below' group relative to the marginal group of 0.03 and nothing for the above group.

4.3.2 Qualifications entered

To study changes in the number of qualifications entered by pupils, we use a shorter dataset with two pre-reform years (2014 and 2015) rather than three because the Wolf reforms led to wholesale changes in the number (and type) of qualifications entered by pupils in 2014 (Burgess & Thomson, 2018). We observe a substantial rise in qualifications entered in the post-reform period by pupils in the 'below' group: relative to marginal pupils, pupils in the 'below' group entered an additional 0.42 of a GCSE (effect size=0.33 SD), while the number of entries from pupils in the 'above' group fell by -0.32 of a GCSE (effect size=-0.25SD). In other words, the numbers of qualifications entered by pupils in the 'marginal and 'below' groups increased by a greater margin than the 'above' group following the introduction of Progress 8. This suggests that pupils in the 'above' group were already entering sets of qualifications well aligned to the new measure. Its introduction provided an incentive to schools to better align the qualifications entered by 'marginal and 'below' pupils to the measure.

4.3.3 Completion

We use equation 2 from section 3.6 to control for the pre-treatment trends shown in Figure 5 and Figure 6 to look at the probability of completing Key Stage 4 at a state-funded mainstream school and the probability of completing Key Stage 4 at the school where pupils started Year 8. Compared to the borderline group, there are slight increases in the odds of completion for below-borderline pupils on both measures following the

introduction of Progress 8. Put another way, borderline pupils are less likely to complete Key Stage 4 in the state-funded mainstream sector following the reform. However, as Figure 5 and Figure 6 indicate, there is a decrease in completion following the reform for all three groups. It is not necessarily the case that Progress 8 is the sole cause of this. Reductions in per-pupil funding (Sibieta et al, 2019) and the 2013 reforms to special educational needs and disabilities may be contributory factors (Thomson, 2019).

Table 5: Other outcomes

	1. 5ACEM (sd units)		2. Mean GCSE (sd units)		3. A8 Entries (post-Wolf)			4. Completion of Year 11 in a mainstream school			5. Completion of Year 11 in the school in which enrolled in Y8		
	b	se	b	se	b	se	Effect Size	b	se	exp(b)	b	se	exp(b)
above	0.224**	0.003	-0.019**	0.002	0.186**	0.005	0.15	0.057	0.032	1.059	0.049*	0.021	1.050
below	-0.249**	0.003	0.012**	0.002	-0.218**	0.007	-0.17	-0.100*	0.030	0.905	-0.046*	0.020	0.956
Interaction of above and reform	**-0.011***	**0.004**	**0.004**	**0.003**	**-0.320****	**0.006**	**-0.25**	**0.046**	**0.035**	**1.047**	**0.035**	**0.024**	**1.035**
Interaction of below and reform	**0.037****	**0.004**	**0.034****	**0.003**	**0.420****	**0.008**	**0.33**	**0.065***	**0.033**	**1.068**	**0.051***	**0.023**	**1.052**
Number of pupils (thousands)		3094		3094		2030			3145			3145	
Number of schools		3165		3165		3158			3330			3330	
R-squared		0.47		0.63		0.32			-			-	

* significant at the 5% level, ** significant at the 0.1% level

Notes

1. Standard errors for columns (1) to (3) are clustered by school*year dummy and columns (4) and (5) by school
2. Columns (1) to (3) are based on equation 1 (see Section 3.5) and columns (4) and (5) on equation 2
3. Pupil covariates are standardized Key Stage 2 score, free school meal eligibility, ethnicity, gender, month of birth, IDACI decile, first language (English/ other) and interactions between each pupil characteristics and standardized Key Stage 2 score. Standardised Key Stage 2 score is fitted as a third-degree polynomial (cubic).

Columns (1) to (3) are based on pupils who reach the end of Key Stage 4; Columns (4) and (5) on pupils observed in Year 8 (see Section 3.2)

4.6 Differential school responses

4.6.1 Different market circumstances

The hypothesis underlying this report is that schools respond to the accountability incentives they are given, and to some degree assign their discretionary resources (for example, their most effective teachers) to the pupils most likely to improve their performance. A school close to the floor standards, for example, would have a stronger incentive pre-reform to focus its attention on borderline pupils and so, a priori, we might expect them to respond to the reform in a different way to higher-performing schools.

We expect that the impact of the reform would be greater in cases where schools had reacted more to the old regime, and this is the basis for most of the cases below. We define a number of school groups, and interact group membership with the difference-in-difference parameters. Results are summarised in Figure 14.

First, we consider schools under strong pressure from being near the floor standard that existed prior to the introduction of Progress 8: at least 40% of pupils achieving five or more A*-C grades at GCSE (or equivalent) including English and maths (5ACEM). For them, the desire to increase the performance of borderline pupils prior to the reforms was likely to be intense. We define this group as schools having performance in the previous year between 35% and 45% 5ACEM. Perhaps the most striking results are that the largest effects can be observed in schools that were close to the floor standard. Once the pressure to focus on borderline pupils was removed, the attainment of the above-borderline and below-borderline groups improved more so than in other schools. This fits our expectations: we would have expected these schools to have been more focused on the borderline group prior to the reform either as a result of external accountability pressures or internal strategic behaviours.

Second, we split by school performance as approximated by a measure of contextual value added (CVA), and interacting the lowest quintile and the highest. The effects of the reform were smaller in schools with high contextual value added. We might surmise that there tended to be less of a focus on borderline pupils prior to the introduction of Progress 8 in these schools.

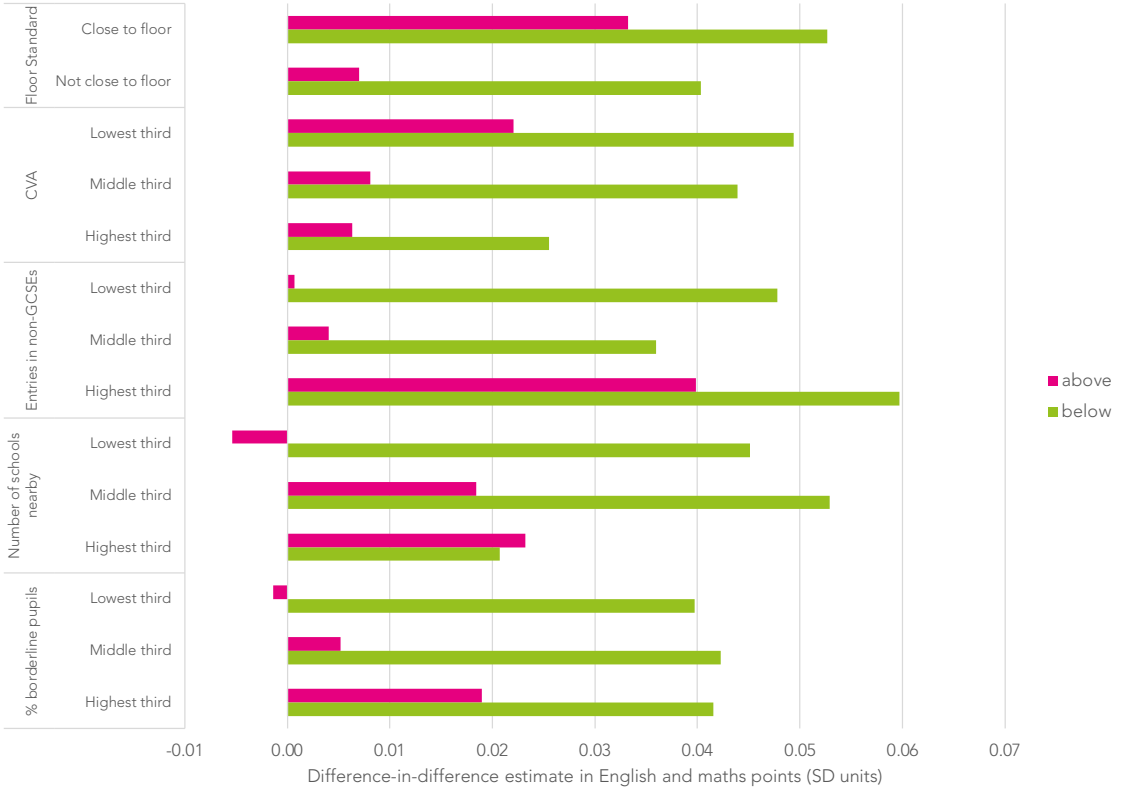
Third, we use a metric that has been taken to characterise strategic behaviour by schools, namely the extent to which they use non-GCSE qualifications. We find that schools making greater use of non-GCSEs also reacted strongly to the removal of the 5ACEM threshold.

Fourth, we consider competitive pressure on schools from the density of alternatives available to parents, measured here by the number of other state-funded mainstream schools within a 3km straight-line distance of the focus school. We know from Burgess et al (2013) that the presence of school performance tables causes schools to focus on and improve their measured performance. Schools for which these competitive forces felt more immediate might be expected to maximise their chances in the market by strongly engaging in prioritising the borderline students. We would therefore expect the removal of the threshold effect to produce bigger changes away from the borderline group in highly competitive areas. Although we see this for the above-borderline group, we do not for the below-borderline group.

Fifthly, we look at variations between schools in the fraction of borderline pupils. We emphasise again that this is an endogenous variable, school performance and admissions will affect this. This might matter for the following reason: schools with just a few borderline pupils would be well placed to channel resources as they could target that quite intensively on the few borderline pupils. A school in which a substantial fraction are borderline however, would it find it much less cost-effective. Finally, we do not see any material differences for below-borderline pupils with respect to the fraction of borderline pupils at a school. This is consistent with column (9) of Table 4. However, there is a slightly larger effect for the above-borderline group. This would be consistent with our expectation as these schools would have previously had the most to gain in terms of published performance indicators by focusing on the borderline group.

Figure 14 summarises the effect of the reform for the above-borderline and below-borderline groups on the different types of schools described above.

Figure 14: Difference in difference estimates interacted with school characteristics, English and maths APS (SD units)



4.6.2 Different school types

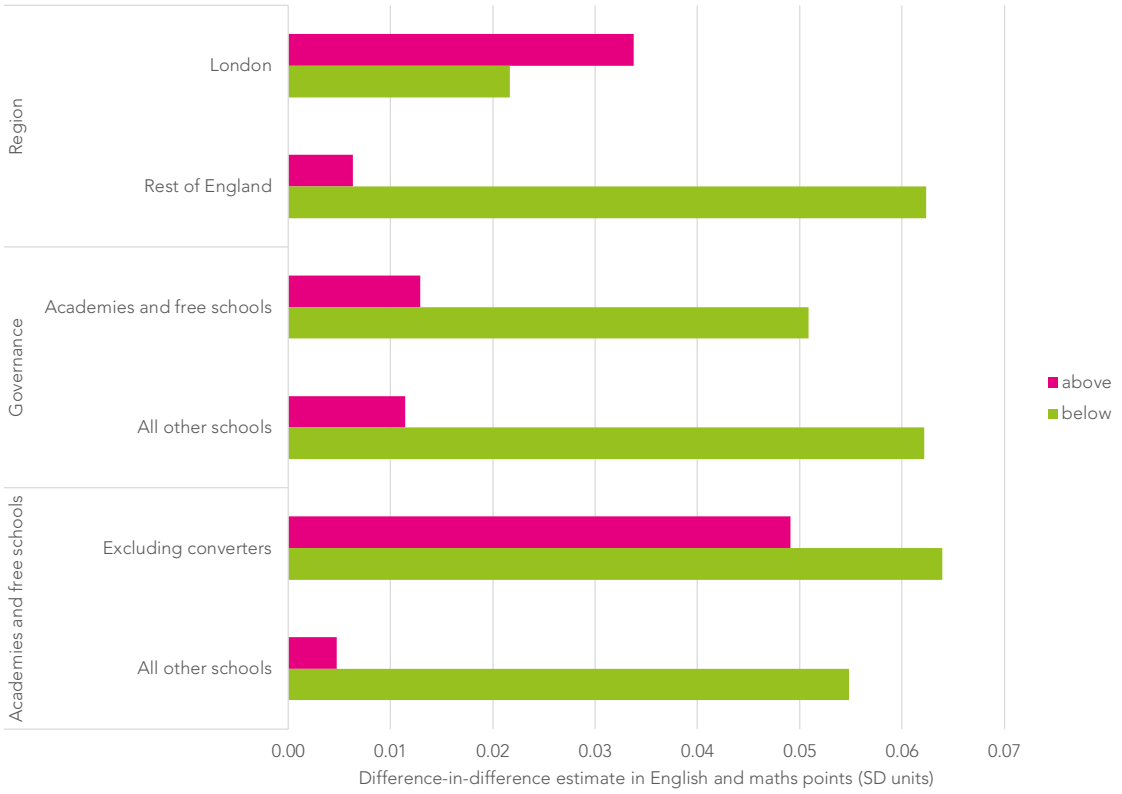
We might expect different school types to respond differently if they have differing levels of autonomy or different pupil intakes (beyond characteristics we control for). These differential responses are summarised in Figure 15.

First, we consider schools in London (Burgess, 2014), where pupils in the above group made post-reform gains relative to borderline pupils although this is accompanied by the below group falling further behind. There are significant differences between schools in London and those elsewhere in the country. Following the introduction of Progress 8, the attainment of above-borderline pupils relative to borderline pupils increased by a greater margin in London than elsewhere. This finding stands out as being very different to all other schools, and as yet remains unexplained in the context of the questions addressed in this paper.

Secondly we assess the impact of the reform in Academies, a new type of state-funded independent schools introduced into the English education system since the turn of the millennium (see Eyles and Machin, 2017 for a fuller discussion). Whether a school is an academy or not is time varying. During the estimation period from 2012 to 2017, the number of academies in the dataset increased from 1,400 to 2,100. When considering all academies, there are no significant differences in post-reform effects. However, there is a significant larger effect for the above-borderline group when converter academies, which by 2017 were the most common type of school, were removed. The remaining set[xvi] of academies consist chiefly of sponsored academies, typically low attaining schools taken from local authority control and handed to non-profit making trusts, and funded by central government (West, 2015). Following the reform, over half of these schools were in the highest third of school for non-GCSE entries shown in Figure 14. The results of both Figure 14 and Figure 15 suggest that these schools responded the most to the accountability reforms. They also tended to have higher proportions of borderline pupils, typically 15% in the pre-reform years compared to 13% in other schools.

Figure 15 summarises the effect of the reform for the above-borderline and below-borderline groups on these different types of schools.

Figure 15: Difference in difference estimates interacted with school type, English and maths APS (SD units)



4.7 Disadvantaged pupils

The estimates set out above describe the causal effect of changes to the accountability system on pupil outcomes. Here we pull out and highlight the effect on disadvantaged pupils. First, we simply take a direct approach and directly estimate the effect of the introduction of Progress 8 on three attainment outcomes for disadvantaged pupils, shown in Table 6. These regressions have the same format and the same control set as the main regressions in Table 3. We focus on the models that include pre-treatment trends as there are small but statistically significant and negative pre-treatment trends for disadvantaged pupils relative to other pupils in all three outcomes.

Table 6: Outcomes for disadvantaged pupils

Specification	Parameter	1. EM points		2. 5ACEM		3. Mean GCSE	
		b	se	b	se	b	se
1. Assuming common trends	fsm6	-0.202**	0.001	-0.168**	0.001	-0.236**	0.001
	Interaction of fsm6 and reform	**0.014****	**0.002**	**0.004**	**0.002**	**0.011****	**0.003**
	Number of pupils (thousands)		3094		3094		3094
	Number of schools		3165		3165		3165
	R-squared		0.67		0.47		0.64
2. With linear trend parameter	fsm6	-0.198**	0.004	-0.124**	0.004	-0.233**	0.004
	Interaction of fsm6 and reform	**0.010***	**0.005**	**0.016****	**0.004**	**0.025****	**0.004**
	Number of pupils (thousands)		3094		3094		3094
	Number of schools		3165		3165		3165
	R-squared		0.66		0.46		0.63

** significant at the 1% level

Notes

1. Standard errors are clustered by school*year dummy
2. Pupil covariates are standardized Key Stage 2 score, disadvantage (FSM6) ethnicity, gender, month of birth, IDACI decile, first language (English/ other) and interactions between each pupil characteristics and standardized Key Stage 2 score
3. Details of the two specifications are given in equations 1 and 2 in Section 3.5

There were slight increases across all three attainment indicators among disadvantaged pupils following the reform, in particular there was an increase in 0.010 SD in EM points. Note that because this is regression-based, it is the impact of disadvantage on test scores, holding constant other factors included in the regression that may co-vary with disadvantage.

Secondly, we provide some insight on the source of that improvement in scores for disadvantaged pupils by using our main results (Table 3, column 4). Using those results, we can simply perform the calculation for the change in the impact of disadvantage due to the reform from the policy treatment effects (the "above*after" and "below*after" coefficients) and the differential membership rates in those two groups of disadvantaged pupils.

We show this calculation in Table 7 below: for disadvantaged pupils (row 2) and non-disadvantaged pupils (row 3), compute the percentage of pupils in the above and below groups and then the difference between the two groups (row 3 – row 2) in row 4. We then multiply these differences by the policy treatment coefficients from Table 3, column 4, shown here in row 1.

Table 7: Impact estimates for disadvantaged pupils based on main results in Table 3, column 4.

Row	Measure	Above	Below
1	Coefficient	0.010	0.057
2	Percentage of not disadvantaged pupils in respective groups:	63%	24%
3	Percentage of Disadvantaged pupils in respective groups:	39%	45%
4	Difference in percentage (row 2 – row 3)	-24%	21%
5	Difference*coefficient (row 4 * row 1)	-0.002	0.012

The predicted overall impact using our model is the sum of the two items in row 5, equal to 0.010, the same as in the "reduced form" estimate in Table 6.

The interpretation that our model brings is that the improvement for disadvantaged pupils mostly arises because the 'below' group sees the largest improvement in scores and disadvantaged pupils are disproportionately found in this group.

5 Behaviour change in schools

The results presented above indicate that Progress 8 brought about a change in attainment for higher and lower attaining pupils relative to pupils working at the C/D

borderline. But they tell us little about the responses made by schools that brought about those changes, save for changing the type of qualifications pupils entered.

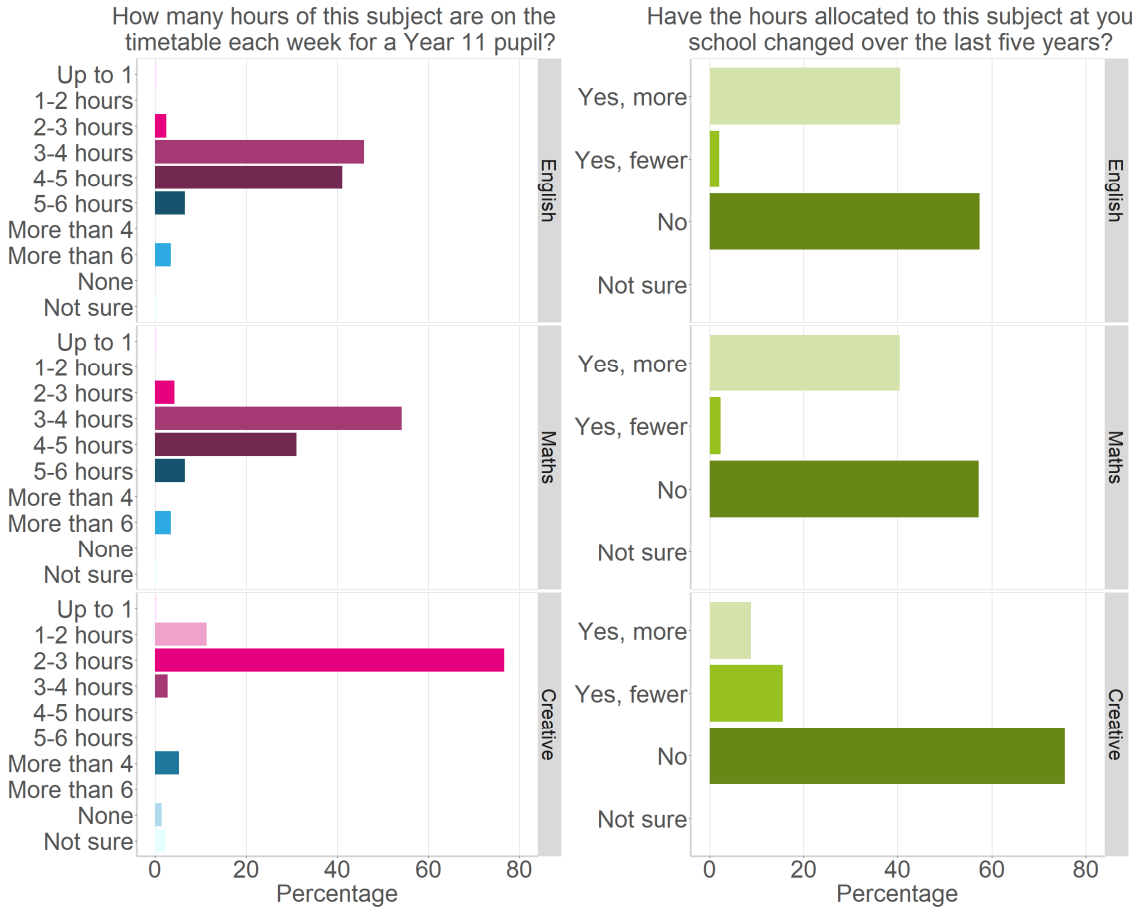
In order to understand a little more about how school behaviour changed in response to Progress 8, we turn to the responses to the online survey we sent to schools described in Section 3.7. We are cautious about reading too much into the results given that we were asking respondents to recall events from four to five years earlier. However, they do reveal something about teachers' post-hoc perceptions of the impact of Progress 8 on the organization of teaching and learning.

Panels summarising the results for each of the four themes of the survey (timetabling, interventions, curriculum structure, teacher allocation) are presented below. We base our results on the 298 schools at which respondents were teaching the year before the introduction of Progress 8.

5.1 Timetabling

Respondents generally reported that there were 3-5 hours allocated on the Year 11 timetable to each of English and maths (Figure 16). Around 40% indicated that this was an increase compared to five years earlier. This is likely to reflect increased content in both subjects. Each creative subject (art and design, drama, music) tended to have 2-3 hours on the Year 11 timetable, with around three quarters of respondents indicating no change compared to five years earlier. It could be the case, however, that some schools are now running fewer options in other subjects in order to accommodate more time for English and maths.

Figure 16: Responses to questions about timetabling



5.2 Interventions

We asked respondents about five groups of pupils for which they may run intervention sessions in Year 11:

- Pupils judged by teachers to be falling behind
- Pupils at the 4/3 (formerly C/D) borderline
- Pupils at the 5/4 borderline
- Disadvantaged pupils
- Pupils falling behind target grades

We firstly asked respondents to rank the above five groups in order of priority for intervention sessions. We then asked if they thought that the school now ran more sessions for each group compared to five years earlier with the exception of the 5/4 borderline group, which was not previously identifiable.

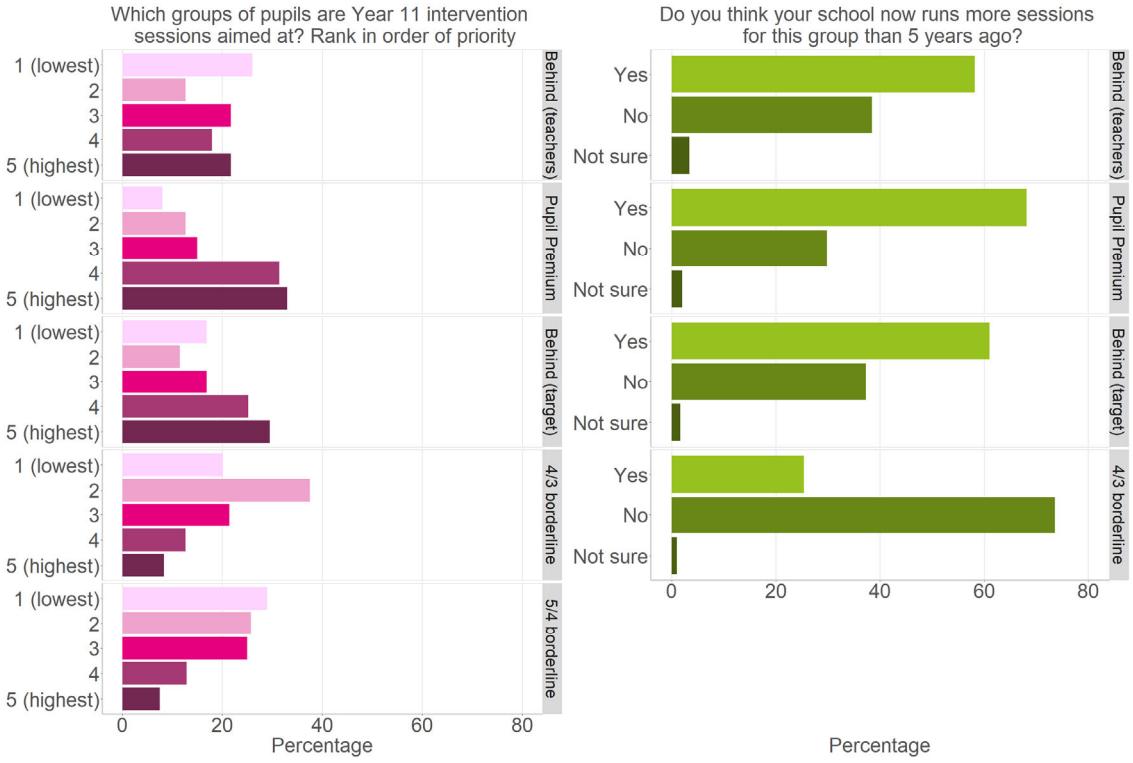
There was a broad spread of results to the question about current priority though a slightly higher percentage of respondents selected disadvantaged pupils and pupils falling behind

target grades as their highest priority (Figure 17). The level of priority given to disadvantaged pupils is likely to be driven by the demands of the Pupil Premium, the government's policy to improve attainment for that group (Department for Education, 2020).

We also asked respondents whether there were any other groups of pupils at which intervention sessions were aimed. 28 (9%) respondents indicated that sessions were provided for pupils with higher levels of prior attainment, 23 (8%) for pupils with special educational needs or disabilities and 16 (5%) for boys.

On the whole, the majority of respondents felt that their schools were now offering more intervention sessions than five years earlier for pupils falling behind target grades (or by teacher judgment) and for disadvantaged pupils. By contrast, the majority felt that they were not offering more sessions for pupils at the 4/3 borderline.

Figure 17: Responses to questions about interventions



5.3 Curriculum structure

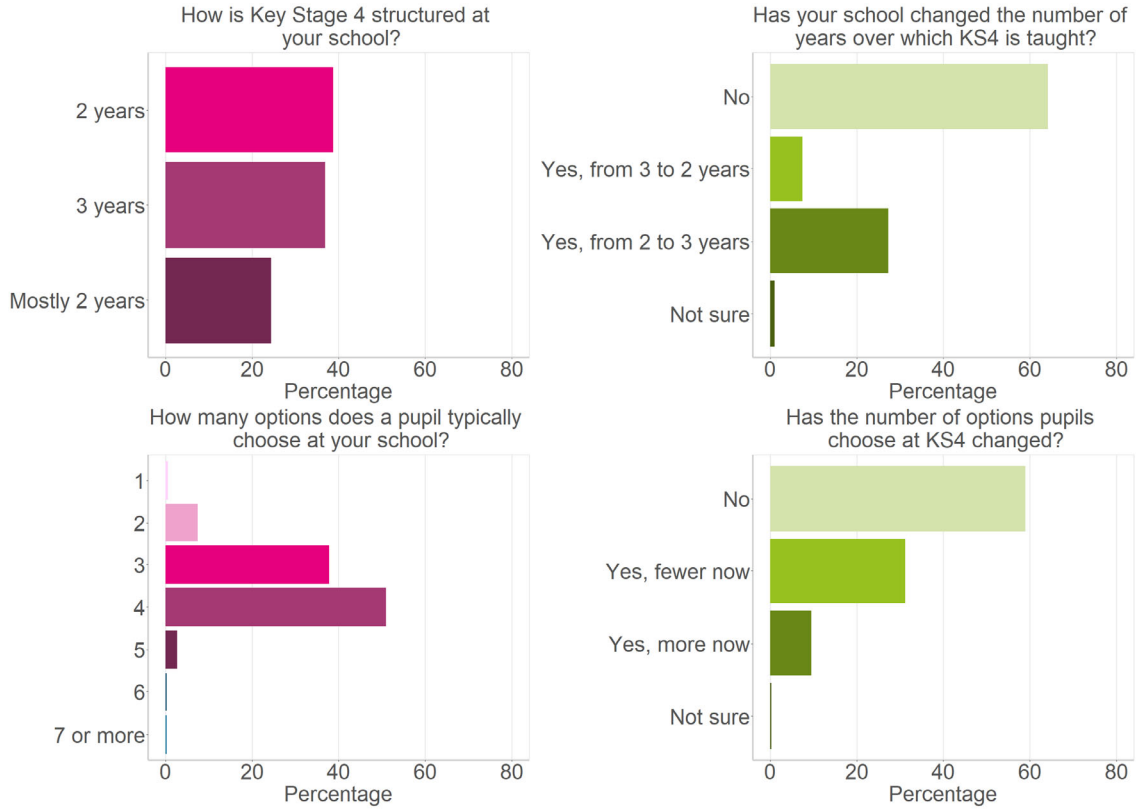
We asked two questions about curriculum structure: the number of years over which Key Stage 4 is taught and the number of optional subjects typically taken by pupils.

A slightly higher percentage of respondents (39%) reported that Key Stage 4 was taught over two years than over three years (37%). The remaining 24% followed a mixed model, with some subjects taught over three years (Figure 18). Among respondents who were working at the school prior to the introduction of Progress 8, 27% reported that their

school had changed to a three-year Key Stage 4 in the last five years. Far fewer (8%) had switched from three-year to two-year. A lively debate about the merits of two-year versus three-year Key Stage 4 was taking place at the time of writing (Harford, 2020).

51% of respondents reported that pupils typically took four optional subjects and a further 38% that they took three. 31% of respondents who were at the same school five years earlier reported that pupils now took fewer options.

Figure 18: Responses to questions about curriculum structure

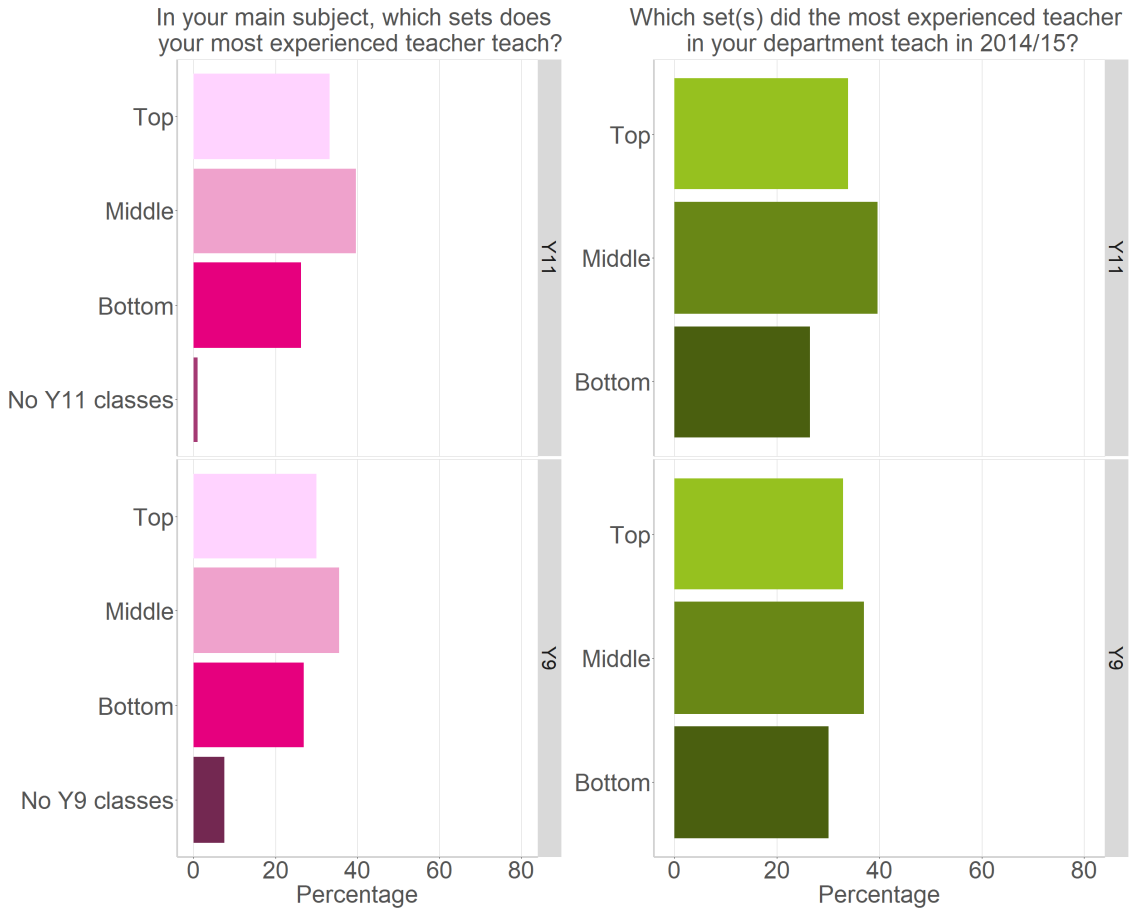


5.4 Teacher allocation

Finally, we asked some questions about the allocation of teachers to sets in Year 11 and Year 9 in the subject taught by respondents if they taught a subject which put pupils into sets.

On the whole, it was slightly more likely for respondents to report that the most experienced teacher in their subject taught middle sets. The situation five years earlier was broadly similar (Figure 19).

Figure 19: Responses to questions about sets taught by the most experienced teacher in a department



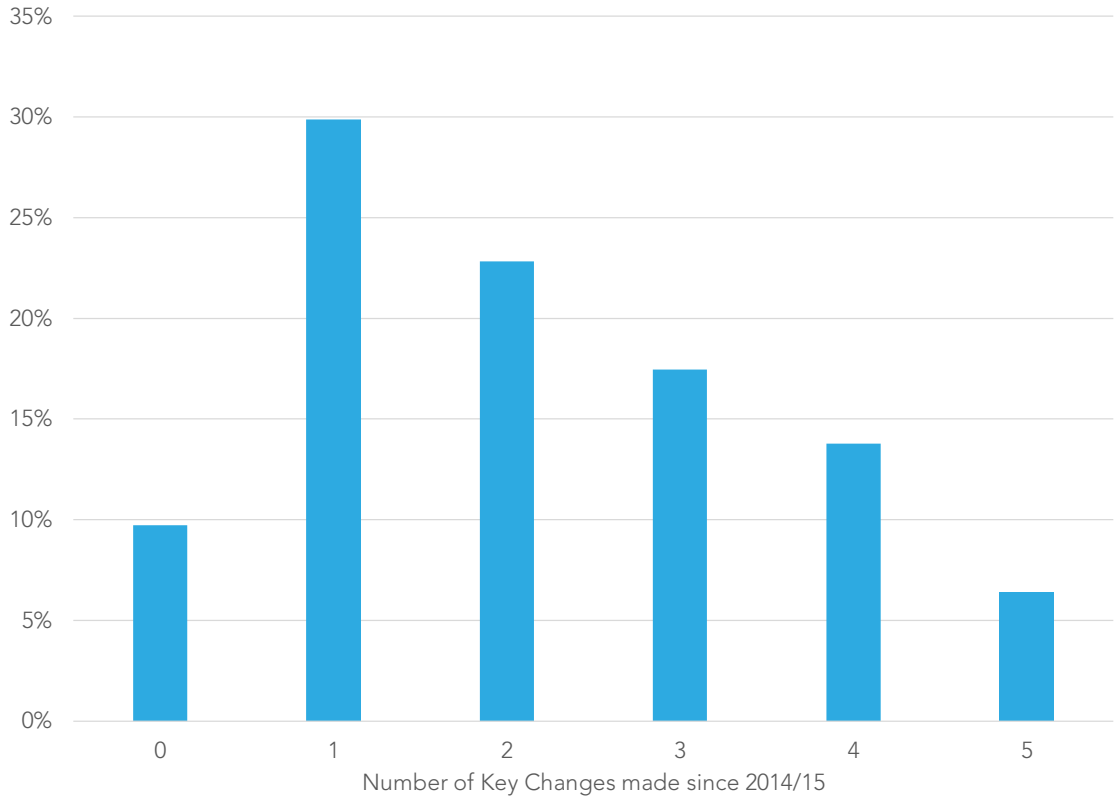
5.5 Combinations of changes

The survey suggests that schools differed in their responses to Progress 8. Respondents to the survey also varied in their views towards Progress 8 as revealed by responses to open comments, with 10 respondents firmly stating that nothing had changed in their school.

But for others, it led to some change. However, the numbers and types of changes varied by school. In the chart below we look at the five major changes highlighted in the previous section: increased time for English, increased time for maths, more intervention sessions for pupils falling behind, fewer options and changes to the length of Key Stage 4.

10% of respondents made none of these changes and 6% made all five (Figure 20).

Figure 20: Number of Key Changes made (n=298)



Of the schools which increased the amount of time available to teach GCSE maths, 86% also increased the amount of time available to teach English. They were also more than twice as likely (46%) as other schools (21%) to report that they had reduced the number of options available.

5.6 The association between behaviour change and the effects of the Progress 8 reforms

We examined how our main results presented in Table 3 vary with respect to the school behaviour change revealed by the survey. We do not make any claims of causality here; we observe whether the effects of the reform on pupils in the above-borderline and below-borderline groups vary with respect to school behaviour change. The reasons for our caution are that 1) we cannot be sure exactly when school behaviour changed; 2) behaviour may have changed in response to other contemporaneous events, such as reduced per-pupil funding and GCSE reforms and 3) we are relying on respondents recalling events from five years earlier.

We firstly run our main specification for the subset of 298 schools that 1) responded to the survey and 2) were able to provide responses to questions about the organisation of teaching and learning five years earlier. This yields similar, though fractionally smaller, effects of the reform on above-borderline (0.004 SD) and below-borderline pupils (0.051 SD). The effect for the above-borderline group is not statistically significant. Because we

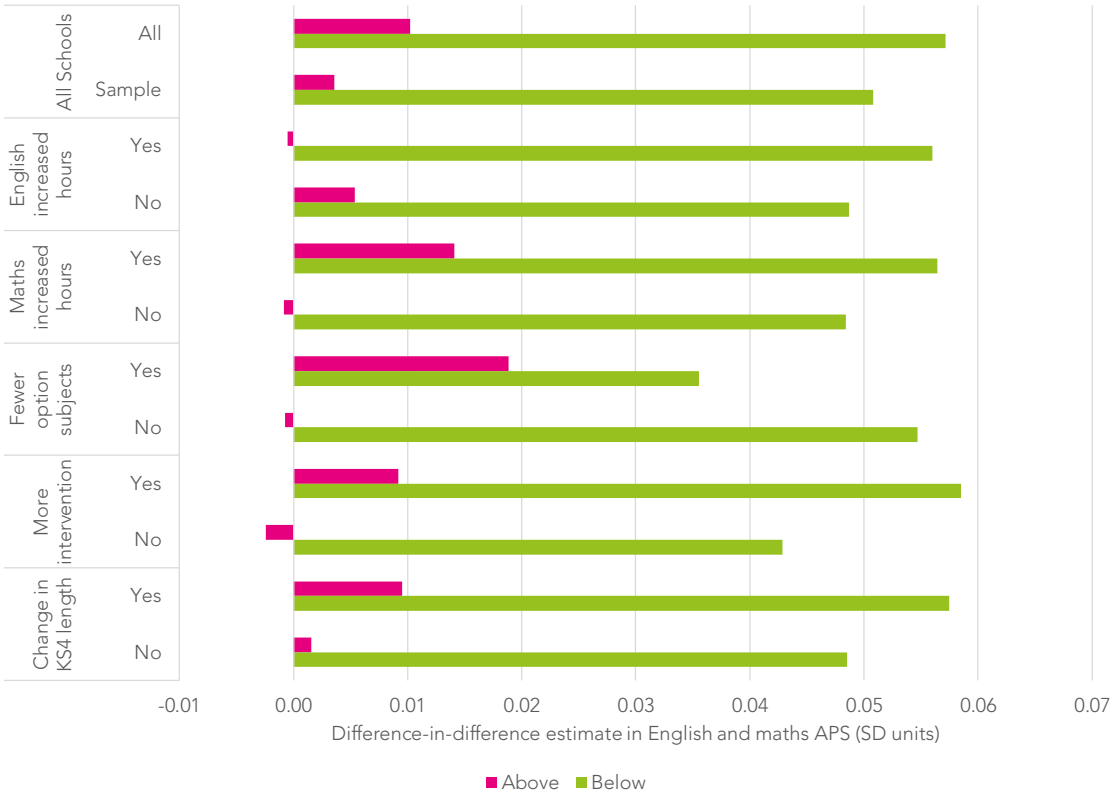
are working with a subset of the sample, the standard errors are larger, in other words our estimates are less precise.

As in Sections 4.4 and 4.5, we then interact these effects with five key behaviour change indicators revealed by the survey:

- Increasing the number of hours on the Year 11 timetable for English
- Increasing the number of hours on the Year 11 timetable for maths
- Reducing the number of option subjects at Key Stage 4
- Providing more intervention sessions aimed at pupils at risk of falling behind
- Changing the length of Key Stage 4

These interaction effects are shown in Figure 21. None of them are statistically significant, however a number of them point in the direction we might expect. Larger effects for both the above-borderline and below-borderline groups can be observed for schools which ran more intervention sessions for pupils who were falling behind, those that increased timetabled hours for maths and those which changed the length of Key Stage 4. An increase in timetabled hours for English is associated with a greater effect of the reform for below-borderline pupils but not above-borderline pupils. Fewer optional subjects is associated with the contrary, a larger effect for the above-borderline group but a smaller effect for the below-borderline group.

Figure 21: Difference in difference estimates interacted with school behaviour change indicators, English and maths APS (SD units)



6 Conclusion

Understanding the appropriate parameters of school accountability continues to be an important subject for research internationally. For example, Dee (2020) looks back on 30 years of school accountability in the US, and analyses the likely impact of the latest federal framework for this. International comparative research using the PISA data, Bergbauer et al (2019), shows the effects of universal standardised testing in 59 countries. There is also attention on mechanisms for these effects: for example, Rouse et al (2013) study how schools in Florida respond to accountability pressure.

We contribute to this evidence by using six years of attainment data on secondary schools in England to explore schools' reactions to significant changes to their accountability framework. The results are consistent with the view that some schools had reacted to the previous regime of high implicit incentives for the test scores of a particular group of students. Once that incentive was removed, that specific group appear to make less relative progress. The effects are not trivial: our headline findings show a post-reform gain of 0.01SD for the above-borderline group and 0.06SD for the below-borderline group.

We have been cautious in presenting these results noting the issue of trends subsequent to announcement but before implementation. We judge the results to be supportive of the hypothesis but not clinching. These results, however, do appear to be robust to a variety of other specification tests.

These results have a bearing on the test score gap between disadvantaged pupils and their peers. Our findings show a post-reform improvement of between 0.01 and 0.02 SD for disadvantaged pupils, which can be decomposed in terms of the accountability-relevant groups.

Schools' responses to the introduction of Progress 8 were varied, although it is difficult to disentangle specific responses to Progress 8 from responses to other changes that happened around the same time, such as reforms to GCSEs. But the results of our survey suggest a general shift away from running intervention sessions aimed specifically at borderline pupils towards pupils judged to be falling behind.

This analysis has messages for policy-makers. First, and bearing in mind the caveats noted above, our results suggest that the introduction of Progress 8 had the intended effect of shifting schools' focus away from students who were marginal to the previous accountability threshold. The effect is not trivial but nor is it a dramatic change. In that sense, the policy "worked". Second, this reinforces the view that accountability measures are an effective policy tool. They do not impinge directly on schools' operational autonomy, unlike explicit Ministerial directives, but they do adjust the incentive structure that schools face. This research shows that this can be effective in changing behaviour. The setting, and occasional re-setting, of the accountability framework seems an appropriate role for Government – it is the practical expression of its view of what society deems valuable in education, of what schools 'ought' to do. Problems arise if the framework is changed very frequently so that schools do not have a stable environment for planning. Problems can also arise if different parts of schools' incentives pull in different directions, and this is the third and final policy message from this study. The previous accountability regime was based on the 5ACEM threshold, so schools were strongly incentivised to maximise the fraction of their pupils that achieved this. This drive meshed well with the goal of the typical pupil because for her passing that threshold was key to access to higher or further education and to the job market. Schools could allocate their resources knowing that the goal of doing well by their pupils and the goal of doing well on the performance metrics were closely aligned. In the new regime, currently, that is less true. Access to higher education and to jobs is still to an extent dominated by the 5ACEM threshold, and this may mean that schools are partially conflicted, and that a goal for the school of keeping the 5ACEM "pass rate" high is still important to them. This may partly explain why the impact of the reform on test scores was rather modest. It may be that the labour market and HE admissions will respond and place more emphasis on P8 scores, or it may be that these two goals for schools will remain in tension.

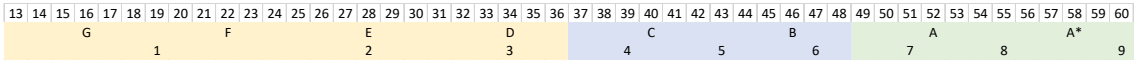
The change to the accountability framework was far-reaching and had other implications beyond simply test scores. The machinery of school accountability also incentivises schools to enter pupils in particular qualifications. Just as the Government response to the Wolf Review had done two years earlier (Burgess and Thomson, 2019), the introduction of Progress 8 led to large changes in the types of qualification for which pupils were entered

as schools increasingly began to fill the eight qualification 'slots' available in the new accountability measures. Pupils in the 'below' group filled on average an additional 0.42 slots compared to borderline pupils who themselves filled on average an additional 0.32 slots compared to pupils in the 'above' group. In most cases, this was a result of switching away from other types of qualification that were not eligible for inclusion in the accountability measures. A lively debate persists about the merits of this (Richmond, 2019).

Finally, the results are consistent with previous international research into school accountability, particularly that of Reback (2008) in Texas. To the extent that the analysis supports the hypothesis, it is clear that schools significantly adjusted their resource allocation strategies in response to the new accountability framework. It is not the case, that schools in general simply "try to do what's best" for their pupils, but respond to the incentive structure they are given. Whether this is seen as positive or negative depends on the social value placed on the educational achievements of the borderline group relative to other groups.

Appendix A: Points score conversion for reformed GCSE

Reformed GCSEs (graded 9-1) were first awarded in English and maths in 2017. In the absence of any existing conversion of these grades into the points scale used by the Department for Education between 2005/06 and 2015/16, we assign points to 9-1 grades as follows:



This yields a reasonably similar means and standard deviations of points for 2017 compared to 2016 for all three parts of the distribution (Table A1).

Table A1: Average point scores for English and maths, 2016 and 2017

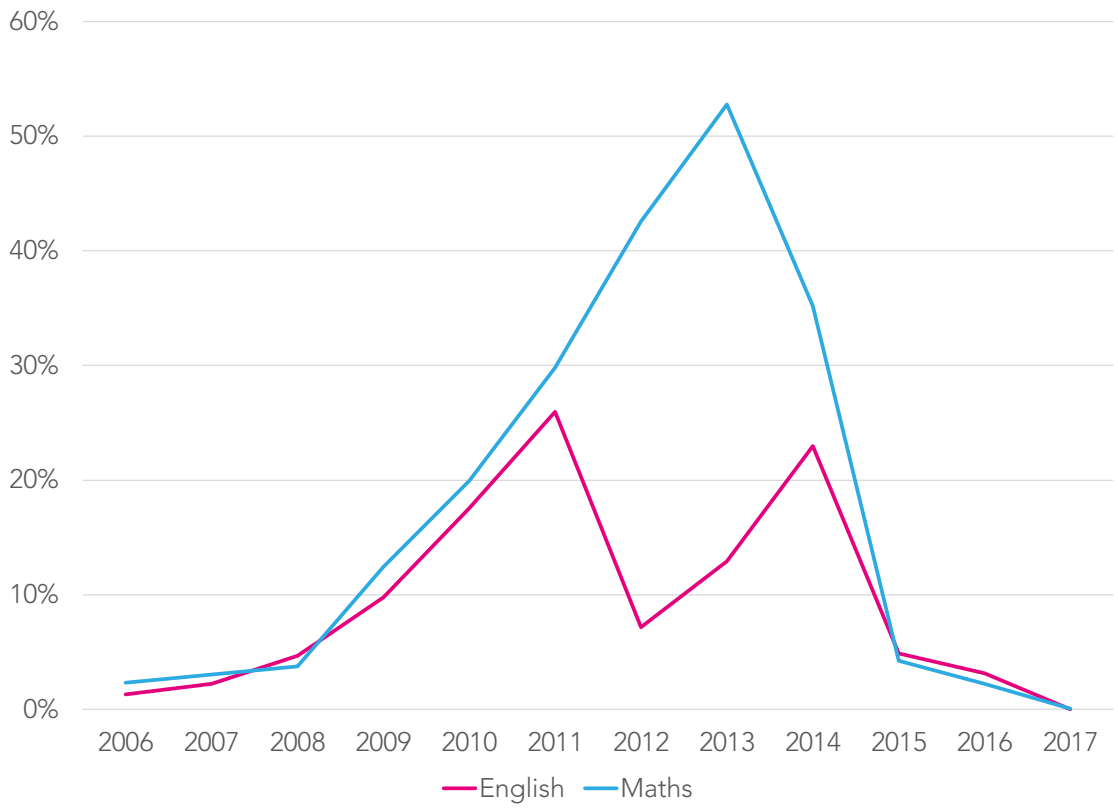
		Average points score		Standard Deviation	
		2016	2017	2016	2017
English	D-G	31.0	31.0	4.7	4.7
	B-C	42.6	43.0	3.0	3.2
	A*-A	53.5	53.6	3.1	3.3
Maths	D-G	28.8	28.8	6.4	5.8
	B-C	42.4	42.2	2.9	3.1
	A*-A	54.4	54.1	4.5	4.3

Appendix B: Multiple entry in GCSE English and Maths

During the period we analyse, 2011/12 to 2016/17, schools responded to changing incentives offered by Performance Tables. These affect comparisons of school performance over time. We analyse the impact of a large change, the introduction of Progress 8 in 2015/16, on pupil attainment in GCSE English and maths. However, there were a number of changes in the immediately preceding as we set out in Section 2 which affect the stability of the outcome measure time-series.

One such example we illustrate here is the practice of entering pupils more than once in GCSE English or GCSE maths. Figure B1 shows the long-term trend in multiple entry. This was particular prevalent in maths, reaching a peak in 2012/13.

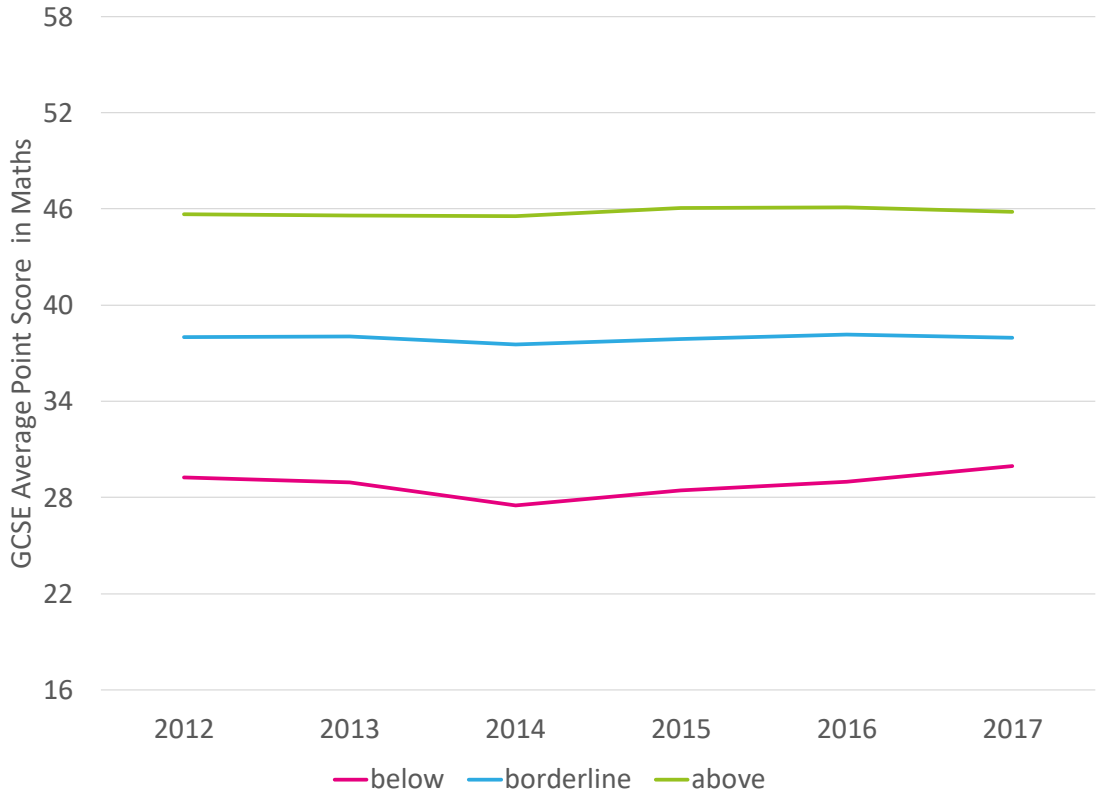
Figure B1:Percentage of pupils entered more than once for GCSE English and GCSE maths, 2006 to 2017



The practice was curbed from 2013/14 onwards as part of a number of changes made to Performance Tables, including the Coalition government's response to the Wolf reforms (Department for Education, 2015b). Up until that year, a pupil's best result in a subject would count towards Performance Tables. This was thought to encourage schools to enter pupils repeatedly in the hope of 'banking' at least grade C (Department for Education, 2012). From 2013/14 onwards, only a pupil's first result would be counted.

Figure B2 shows the average point score in maths during the period we analyse. Pupils have been divided into the three groups relative to their probability of achieving 5 or more A*-C grades including English and maths (Section 3.3). The changes to rules around multiple entry coincide with a dip in attainment for the 'below-borderline' group and, to lesser extent, the borderline group.

Figure B2: Average point score in GCSE maths by pupil group, 2011/12 to 2016/17



The practice of multiple entry affects the time-series in the primary outcome measure we use in our analysis, the average point score in GCSE English and maths points score. We can get a handle on its impact by calculating the average point score in GCSE maths for the 2012/13 had the 2013/14 Performance Tables rules been implemented earlier. A necessary caveat here is that different decisions might have been taken for the 2012/13 cohort if schools had been aware of rule changes then. Averages for pupils in each of the three groups are shown in Table B1.

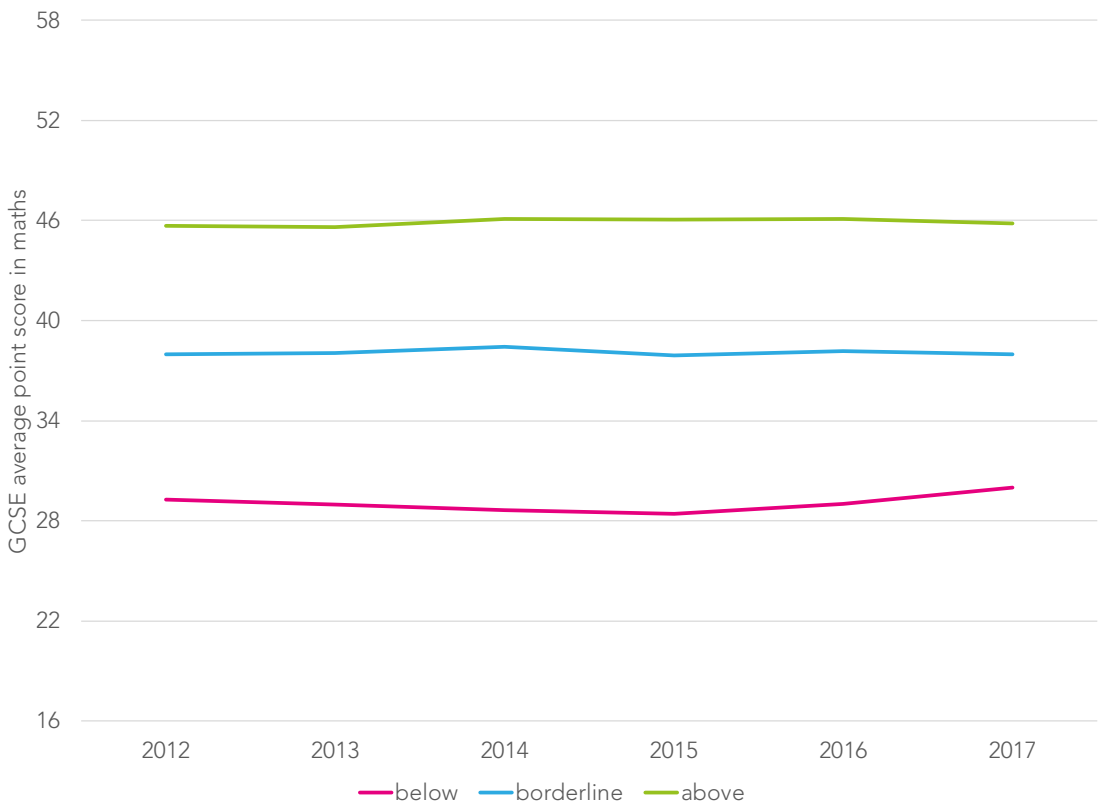
Table B1: GCSE average point score in maths, 2012/13 by pupil group

Pupil group	2014 basis	2013 basis	Difference
Below	27.82	28.96	-1.14
Borderline	37.17	38.04	-0.87
Above	45.06	45.60	-0.54

The application of the 2014 rules appears to have had the greatest impact on the below group. Compared to the borderline group, their points score fell by an additional 0.27 points (-1.14-0.87). By contrast, the above group improved relative to the borderline group with the 2014 rules applied. These differences are equivalent to around 0.02 of a standard deviation.

If we assume that average point scores in GCSE maths were affected by a similar margin in 2013/14, we can add the absolute values of the differences in Table B1 to the values for 2013/14 in Figure B2. The effect of this is shown in Figure B3. This results in a smoother line for the period 2011/12 to 2014/15 for the 'below' group.

Figure B3: Average point score in GCSE maths by pupil group, 2011/12 to 2016/17 (with 2013/14 adjustment)



References

Allen, R. (2015) *Opting into 2015 Progress 8 would have been an easier route to avoid floor standards for many secondary schools - FFT Education Datalab*, *FFT Education Datalab blog*. Available at: https://ffteducationdatalab.org.uk/2015/08/opting-into-2015-progress-8-would-have-been-an-easier-route-to-avoid-floor-standards-for-many-secondary-schools/ (Accessed: 26 November 2020).

Angrist, J. D. and Pischke, J. S. (2014) *Mastering 'metrics: The path from cause to effect*. doi: 10.5860/choice.189854.

Astle, J., Bryant, S. and Hotham, C. (2011) 'School choice and accountability: putting parents in charge'. The British Library. Available at: https://www.bl.uk/collection-items/school-choice-and-accountability-putting-parents-in-charge (Accessed: 26 November 2020).

Baird, J.-A. *et al.* (2019) *Examination Reform: Impact of Linear and Modular Examinations at GCSE*.

Benton, T. (2016) *Comparable Outcomes: Scourge or Scapegoat?* Available at: http://www.cambridgeassessment.org.uk/ (Accessed: 26 November 2020).

Benton, Tom; Sutch, T. (2014) *Analysis of use of Key Stage 2 data in GCSE predictions, Ofqual/14/5471*. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/429074/2014-06-16-analysis-of-use-of-key-stage-2-data-in-gcse-predictions.pdf (Accessed: 26 November 2020).

Bergbauer, A.B., E. A. Hanushek, and L. Woessmann (2019) *Testing*. NBER Working Paper no. 24836 (revised). Available at: https://www.nber.org/system/files/working_papers/w24836/w24836.pdf (Accessed: 26 November 2020)

Burgess, S. (2014) *Understanding the success of London's schools*, *Cmpo*. Available at http://www.bris.ac.uk/media-library/sites/cmpo/migrated/documents/wp333.pdf (Accessed: 26 November 2020).

Burgess, S., Greaves, E. and Vignoles, A. (2019) 'School choice in England: evidence from national administrative data', *Oxford Review of Education*. Routledge, 45(5), pp. 690–710. doi: 10.1080/03054985.2019.1604332.

Burgess, S. *et al.* (2005) *Who wins and who loses from school accountability? The distribution of educational gain in English secondary schools*. Available at: www.bris.ac.uk/Depts/CMPO/ (Accessed: 26 November 2020).

Burgess, S. and Thomson, D. (2019) 'The impact of the Wolf reforms on education outcomes for lower-attaining pupils', *British Educational Research Journal*. doi: 10.1002/berj.3515.

Burgess, S. and Thomson, D. (2013) *Key Stage 4 Accountability: Progress Measure and Intervention Trigger*. Available at: www.cubec.org. (Accessed: 26 November 2020).

Burgess, S., Wilson, D. and Worth, J. (2013) 'A natural experiment in school accountability: The impact of school performance information on pupil progress', *Journal of Public Economics*. doi: 10.1016/j.jpubeco.2013.06.005.

Cresswell, M. J. (1996) 'Defining, setting and maintaining standards in curriculum-embedded examinations: Judgmental and statistical approaches', in Goldstein, H. and Lewis, T. (eds) *Assessment: Problems, developments and statistical issues*, pp. 57–84.

Dee, T. (2020) *Learning from the Past: School Accountability before ESSA. A Background Paper for the Hoover Education Success Initiative*. https://www.hoover.org/research/learning-past-school-accountability-essa (Accessed: 26 November 2020).

Department for Children, Schools and Families (2009) *[ARCHIVED CONTENT] Department for Children, Schools and Families : National Challenge*. Available at: https://webarchive.nationalarchives.gov.uk/20090320074819/http://www.dcsf.gov.uk/nationalchallenge/ (Accessed: 26 November 2020).

Department for Education (2010) *The Importance of Teaching*. Available at: http://www.official-documents.gov.uk/ (Accessed: 26 November 2020).

Department for Education (2012) *GCSE early entry: Ofsted asked to discourage a "damaging trend" - GOV.UK, Webpage*. Available at: https://www.gov.uk/government/news/gcse-early-entry-ofsted-asked-to-discourage-a-damaging-trend (Accessed: 26 November 2020).

Department for Education (2013) *Reforming the accountability system for secondary schools Government response to the February to May 2013 consultation on secondary school accountability*. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/249893/Consultation_response_Secondary_School_Accountability_Consultation_14-Oct-13_v3.pdf (Accessed: 26 November 2020).

Department for Education (2015a) '2015 School and College Performance Tables Statement of Intent', (July).Department for Education and Skills (2005) *14-19 Education and Skills White Paper*.

Department for Education (2015b) *Statistical First Release Revised GCSE and equivalents results in England, 2013 to 2014*

Department for Education (2015c) *Hundreds of 'coasting' schools to be transformed - GOV.UK*. Available at: https://www.gov.uk/government/news/hundreds-of-coasting-schools-to-be-transformed (Accessed: 26 November 2020).

Department for Education (2019a) *Education Secretary confirms plans to simplify school accountability - GOV.UK*. Available at: https://www.gov.uk/government/news/education-secretary-confirms-plans-to-simplify-school-accountability (Accessed: 26 November 2020).

Department for Education (2019b) '2019 School and College Performance Tables Statement of Intent', (July).

Department for Education (2020) *Pupil premium policy paper. Webpage. Available at:* https://www.gov.uk/government/publications/pupil-premium/pupil-premium (Accessed: 26 November 2020).

Education committee (2013) *HC 204 [incorporating HC 588-i, 588-ii and 588-iii] 2012 GCSE English results First Report of Session 2013-14 Report, together with formal minutes, oral and written evidence The Education Committee*. Available at: www.parliament.uk (Accessed: 26 November 2020).

Eyles, A. and Machin, S. (2019) 'The Introduction of Academy Schools to England's Education', *Journal of the European Economic Association*. Oxford Academic, 17(4), pp. 1107–1146. doi: 10.1093/jeea/jvy021.

Fischer Family Trust (2020) *Target setting - FFT*, *Webpage*. Available at: https://fft.org.uk/fft/target-setting/ (Accessed: 26 November 2020).

Foley, B. and Goldstein, H. (2012) 'Measuring success: League tables in the public sector', pp. 1–80.

Goldstein, H. and Lewis, T. (1996) 'Assessment: Problems, developments and statistical issues', *Wiley Series in Probability and Statistics*.

Harford, S. (2020) *Making curriculum decisions in the best interests of pupils. Webpage.* Available at https://educationinspection.blog.gov.uk/2020/01/09/making-curriculum-decisions-in-the-best-interests-of-children/ (Accessed 26 November 2020).

Harrison, N., James, D. and Last, K. (2015) 'Don't know what you've got 'til it's gone? Skills-led qualifications, secondary school attainment and policy choices', *Research Papers in Education*. Routledge, 30(5), pp. 585–608. doi: 10.1080/02671522.2014.1002526.

Henshaw, P. (2017) *Does Progress 8 increase the incentives to 'manage-out' pupils? Webpage*. Available at https://www.sec-ed.co.uk/news/does-progress-8-increase-the-incentives-to-manage-out-pupils/ (Accessed: 26 November 2020).

Ingram, J. *et al.* (2018) 'Playing the system: incentives to "game" and educational ethics in school examination entry policies in England', *Oxford Review of Education*. Routledge, 44(5), pp. 545–562. doi: 10.1080/03054985.2018.1496906.

Jenkins, A., Levacic, R. & Vignoles, A. (2005). Estimating the Relationship between School Resources and Pupil Attainment at GCSE. Department for Education and Skills, Research Report RR727.

Leckie, G. and Goldstein, H. (2017) 'The evolution of school league tables in England 1992-2016: "Contextual value-added", "expected progress" and "progress 8"', *British Educational Research Journal*. Blackwell Publishing Ltd, 43(2), pp. 193–212. doi: 10.1002/berj.3264.

Muriel, A. and Smith, J. (2011) 'On educational performance measures', *Fiscal Studies*, pp. 187–206. doi: 10.1111/j.1475-5890.2011.00132.x.

Nye, P. (2017) *Who's left: Will Progress 8 reduce incentives to lose low-attaining pupils? - FFT Education Datalab*, *FFT Education Datalab blog*. Available at: https://ffteducationdatalab.org.uk/2017/01/whos-left-will-progress-8-reduce-incentives-to-lose-low-attaining-pupils/ (Accessed: 26 November 2020).

Ofqual (2011) *GCSEs and A Levels in Summer 2012 Our approach to setting and maintaining standards*.

Ofqual (2016) *An investigation into the 'Sawtooth Effect' in GCSE and AS / A level assessments*. Available at: https://www.gov.uk/government/publications/investigation-into-the-sawtooth-effect-in-gcses-as-and-a-levels (Accessed: 26 November 2020).

Ofsted (1999) *Education Inspection Framework*. Available at: https://www.gov.uk/government/consultations/education-inspection-framework-2019-inspecting-the-substance-of-education/education-inspection-framework-2019-inspecting-the-substance-of-education (Accessed: 26 November 2020).

Ofsted (2008) 'Using data, improving schools', Available at: http://www.ofsted.gov.uk/content/download/6946/71297/file/using data, improving schools.pdf (Accessed: 26 November 2020).

Ofsted (2009) *Attainment: Supplementary guidance for Section 5 inspections*, *Supplementary Guidance and Resources*. Available at: https://web.archive.org/web/20090904174153/http://www.ofsted.gov.uk/content/download/9921/114822/file/Supplementary guidance and resources.zip (Accessed: 26 November 2020).

Ofsted (2013) 'Schools' use of early entry to GCSE examinations', (March), pp. 11–16. Available at: www.nationalarchives.gov.uk/doc/open-government-licence/, (Accessed: 26 November 2020).

Ofsted (2019) 'Education inspection framework 2019: inspecting the substance of education'. Available at: https://www.gov.uk/government/consultations/education-inspection-framework-2019-inspecting-the-substance-of-education/education-inspection-framework-2019-inspecting-the-substance-of-education (Accessed: 8 July 2020).

Osborne, K. (2003) 'Pupil number projections and school place planning in LEAs'. Education-line.

Reback, R. (2008) 'Teaching to the rating: School accountability and the distribution of student achievement', *Journal of Public Economics*. doi: 10.1016/j.jpubeco.2007.05.003.

Revell, P. (2001) *Thomas Telford's 100% GCSE record | Education | The Guardian*, *The Guardian*. Available at: https://www.theguardian.com/education/2001/aug/21/schools.gcses (Accessed: 26 November 2020).

Richmond, T. (2019) *A Step Baccwards: Analysing the impact of the 'English Baccalaureate' performance measure*. Available at: www.edsk.org (Accessed: 26 November 2020).

Rouse, C.E., Hannaway, J., Goldhaber, D., and Figlio, D. (2013) Feeling the Florida Heat? How Low-Performing Schools Respond to Voucher and Accountability Pressure. *American Economic Journal: Economic Policy*, 5 (2): 251-81.

Sibieta, L., Farquharson, C. and Britton, J. (2019) *2019 annual report on education spending in England*. The IFS. doi: 10.1920/re.ifs.2019.0162.

Simmons, J. P., Nelson, L. D. and Simonsohn, U. (2011) 'False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant', *Psychological Science*. SAGE Publications Inc., 22(11), pp. 1359–1366. doi: 10.1177/0956797611417632.

Thomson, D. (2019) *A data history of permanent exclusions and school moves - FFT Education Datalab*, *FFT Education Datalab blog*. Available at: https://ffteducationdatalab.org.uk/2019/05/a-data-history-of-permanent-exclusions-and-school-moves/ (Accessed: 26 November 2020).

Treadaway, M. (2015) *Why measuring pupil progress involves more than taking a straight line*, *FFT Education Datalab blog*. Available at: https://ffteducationdatalab.org.uk/2015/03/why-measuring-pupil-progress-involves-more-than-taking-a-straight-line/ (Accessed: 26 November 2020).

UK Government (2019a) *All schools and colleges in England - GOV.UK - Find and compare schools in England*, *Webpage*. Available at: https://www.compare-school-performance.service.gov.uk/schools-by-type?step=phase&geographic=all&region=0&phase=secondary (Accessed: 26 November 2020).

UK Government (2019b) *Download data - GOV.UK - Find and compare schools in England*, *Webpage*. Available at: https://www.compare-school-performance.service.gov.uk/download-data?currentstep=datatypes&regiontype=all&la=0&downloadYear=2013-2014&datatypes=ks4 (Accessed: 26 November 2020).

UK Government (2020) *Get information about schools*, *Webpage*. Available at: https://get-information-schools.service.gov.uk/ (Accessed: 26 November 2020).

West, A. (2010) 'High stakes testing, accountability, incentives and consequences in English schools', *Policy and Politics*, 38 (1). pp. 23-39. doi: 10.1332/030557309X445591

West, A. (2015) 'Education policy and governance in England under the Coalition government (2010-15): Academies, the pupil premium, and free early education', *London Review of Education*. UCL Institute of Education, 13(2), pp. 21–36. doi: 10.18546/LRE.13.2.03.

Wilson, D., Croxson, B. and Atkinson, A. (2004) *'What Gets Measured Gets Done': Headteachers' Responses to the English Secondary School Performance Management System*. Available at: http://www.dfes.gov.uk/performancetables/index.shtml (Accessed: 26 November 2020).

Wolf, A. (2011) 'Review of Vocational Education – The Wolf Report', (March), p. 196. doi: 10.1037/h0066788.

[i] Unlike Progress 8, which calculates the mean outcome for pupils in each prior attainment band, the median outcome was used in early value added measures.

[ii] Several versions of points scores have been used over the years. The original scale allocated 1 point to grade G at GCSE up to 8 points for grade A* in intervals of 1 point. This was then replaced in 2004 by a version which allocated 16 points to grade G up to 58 points for A* in intervals of 6 points. The original scale was then brought back for Progress 8 in 2016 and altered to accommodate reformed GCSEs graded 9-1 in 2017.

[iii] Geography, history or ancient history.

[iv] It can easily be seen that this is the headline measure as it is the one presented and highlighted on the landing page of the Department for Education's school performance comparison site (UK Government, 2019a).

[v] 327 schools opted in to a pilot of the Progress 8 measure in 2015. For these schools, we still consider 2015 to be a pre-reform year (i.e. $\tau = 0$) since the option to opt-in was only announced in June 2014 when the affected cohorts were midway through Key Stage 4. In other words, any response to the policy change would have only affected one year of pupils' two-year Key Stage 4 programmes. In one of our tests of robustness we remove these schools from our model specification.

[vi] Note that one of the authors is employed by FFT.

[vii] We do not include pupils without KS2 results (for instance those arriving from overseas during their secondary education) in our analysis.

[viii] This was a single GCSE available until 2016. Since 2017, only separate GCSEs in English language and literature have been available.

[ix] Grade A*=58; A=52; B=46; C=40; D=34; E=28; F=22; G=16, U=0. Pupils not entered are assigned 0 points. Equivalent scores for pupils entering AS-levels are also used. A pupil's highest score is used.

[x] The requirement to have 3 other GCSEs graded A*-C was dropped but in practice, it was rare that pupils achieved A*-C in English and maths without achieving 3 A*-C GCSEs (or equivalents) in other subjects. In 2013, 98.8% of pupils who achieved A*-C in English and maths achieved at least 3 other A*-C GCSEs or equivalents.

[xi] Up to and including 2016, schools could enter pupils in a single GCSE of combined language and literature. Scores in this were also doubled for Attainment 8 purposes.

[xii] Any two of biology, chemistry, physics or computing or core and additional science.

[xiii] Note we do not place a lot of weight on the interpretation of the coefficients on these variables. They are only separately identified from the above and below variables (a_{is} and b_{is}) by functional form.

[xiv] In cases where multiple users subscribed on the same day, we randomly selected one.

[xv] We include pupils who subsequently remain in state-funded education but move to other types of schools (special schools and alternative provision) for whom we observe Key Stage 4 outcomes. We also include anyone who moves into independent schools or home education if they are observed to enter qualifications. However, the percentage of pupils for whom we do not observe Key Stage 4 outcomes, increased from 1.5% among the 2013 cohort to 2.7% among the 2017 cohort. This will include pupils who emigrate and those who remain in England but who are not educated in the state-funded system. Unfortunately, we are not able to quantify the size of either group precisely.

[xvi] In the 2017 data, there were 593 sponsored academies, 51 free schools, 38 university technical colleges (UTCs) and 32 studio schools.